Understanding and Predicting the Dynamics, Folding and Binding of Proteins

Erik Pfeiffenberger

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Biomolecular Modelling Laboratory,

The Francis Crick Institute

and

Faculty of Life Sciences,

University College London

January 18, 2018

I, Erik Pfeiffenberger, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

Proteins are involved in all processes of life and their shapes, interactions and functions are governed by physical forces. A model with atomic resolution is pivotal for the understanding of their mechanisms and how mutations perturb these. However, given the large variation of proteins and the limitations of experimental methods, *in-silico* approaches are the only viable solution. Presented here are a number of computational methods to predict their structure and binary interactions with atomic detail. Firstly, a machine-learning method was developed that models the recognition process of protein-protein binding to improve the identification of near-native binding sites. Secondly, a refinement method was developed to improve the structural accuracy of predicted monomers. An intra reside-residue contact map space was defined to perform more directed conformational exploration with metadynamics in order to find solutions that better resemble the native state. This method was extended to perform refinement of pre-docked heterodimers in order to predict the conformational transition from unbound to bound. Here, an inter residue-residue contact map space was defined between the interface of a receptor and a ligand. Following this extensive sampling of protein conformations by simulation, a recurrent neural network was defined and trained to predict the state changes during the sampling such that improved quality conformations can be identified. Finally, extensive *in-silico* biophysical experiments were performed to understand the mechanism of auto-phosphorylation for RET-kinase in wild-type and its deregulation by an oncogenic mutation.

Impact Statement

The work presented here explored new methods for protein structure prediction of monomers and heterodimers. Given the limitations of experimental methods and the slow rate at which new structures are solved make computational approaches a necessity in order to obtain a complete structural knowledge of all proteins and their interactions.

The machine-learning model developed to identify the correct binding site of two proteins could be applied to annotate the human interaction network with structural predictions where experimental data is missing. Furthermore, more reliable prediction of protein-protein interactions (PPI) could be beneficial for rational drug design. The interest of pharmaceutical companies to develop inhibitor drugs for PPIs increased substantially in the last years. The shift from blocking protein activity, such as kinases, to certain protein interactions allows for the development of small molecule drugs for a wider range of drug targets. Thus, prediction methods of PPIs with atomic detail can be helpful for first exploratory rational-drug design and testing.

The thesis presented results on sampling the conformational states of a large number of proteins with molecular dynamics. Subsequently, a recurrent neural network was trained to identify segments in the trajectory that resemble more closely the native state. The results showed that such a task could be learned by a neural network. This opens up the broader question whether such neural networks or similar ones can learn the protein folding funnel? Progress towards this question could have the potential to solve the protein folding problem.

Acknowledgements

It is impossible to thank all family, friends, colleagues, professors and teachers who have enriched and supported me. To those I have omitted, I apologise.

My warmest thank you goes to my parents, Anke Pfeiffenberger and Lutz Pfeiffenberger, from a young age they have nurtured my curiosity for science and the world and supported me throughout the years. For that I am forever thankful.

I want to thank Dr. Paul Bates for number of things. First of all, for believing in me and allowing me to conduct my PhD research in his laboratory. For giving me the space to pursue my own thoughts but also for his advice and guidance when I needed it. The many discussions helped me to grow as a researcher and for that I am indebted.

A special thanks goes to all current and former members of the Biomolecular Modelling Laboratory. For all the unforgettable memories with them, without them the years at the London Research Institute and The Francis Crick Institute would not have been the same. First, I would like to thank Dr. Raphael Chaleil for being an invaluable member of the laboratory and his help and advice. Dr. Tammy Cheng for her useful discussions and the encouragements. Dr. Rudi Agius for introducing me to machine learning and his music that made writing this dissertation more pleasant. Dr. Robert Jenkins for sharing all his math wisdom with me. Dr. Melda Tozluoglu, Dr. Sakshi Gulati, Dr. Mieczyslaw Torchala, Dr. Xiao Fu, Esther Wershof and Tereza Gerguri for the fun time in and outside of the lab.

ACKNOWLEDGEMENTS

My thanks goes to my thesis advisers Prof. Neil Mcdonald and Prof. Nicholas Luscombe for all their advice from the start to the end of my PhD.

I would also like to thank Dr. Ian Moal from the EMBL-EBI, Prof. David Jones from UCL, Prof. Kathleen Steinhofel from KCL, Dr. Cen Wan from the Jones group, Dr. Aylin Cakiroglu from the Luscombe group and Dr. Dean Plumbley from Benevolent AI for their remarks and insightful discussions on my research in machine learning.

Finally, I want to thank my examiners, Prof. Alexandre Bonvin and Prof. Maya Topf for accepting to review my thesis and allowing me to defend it *viva voce*.

Contents

List of Figures

List of Tables

List of Acronyms

AA	Amino Acid
AI	Artifical Intelligence
AIDS	Acquired Immune Deficency Syndrome
AL	Activation Loop
ANN	Artifical Neural Network
ATP	Adenosine TriPhosphate
CAPRI	Critical Assessment of PRedicted Interactions
CASP	Critical Assessment of protein Structure Prediction
CDK5	Cyclin-Dependent Kinase 5
CM	Contact Map
CMS	Contact Map Space
CS	Combined Scoring
CV	Cross Validation (in the context of machine learning)
CV	Collective Variable (in the context of metadynamics)
DNA	DeoxoyriboNucleic Acid
DNN	Deep Neural Network

DR Distance Restraints

EGFR Epidermal Growth Factor Receptor

EM Energy Minimization

ERT Extremley Randomized Trees

FA Factor Analysis

FES Free Energy Surface

FFT Fast Fourier Transformation

FJC Freely-Jointed Chain

FM Free Modelling

FMTC Familial MTC

FN False Negative

FNAT Fraction of NATive contacts

FP False Positive

GDNF Glial Derived Neurothrophic Factor

GDNF GDNF Family Ligands

GDTHA Global Distance Test High Accuracy

GDTTS Global Distance Test Total Score

GRL Glycine Rich Loop

GRU Gated Recurrent Unit

HIV Human Immunodeficiency Virus

IRMSD Interface RMSD

KNN K Nearest Neighbour

KPCA Kernel PCA

LCO-CV Leave-Complex-Out Cross-Validation

LR Logistic Regression

LRMSD Ligand RMSD

MD Molecular Dynamics

MEN2 Multiple Endocrine Neoplasia type 2

MHC Major histocompatibility complex

ML Machine Learning

MT Mutant Type

MTC Medullary Thyroid Carcinoma

NR No Restraints

PCA Principal Component Analysis

PDB Protein Data Bank

PPMCC Pearson Product Momentum Correlation Coefficient

PR Point Restraints

RF Random Forest

RFE Recusive Feature Elimination

RMSD Root Mean Square Deviation

RMSF Root Mean Square Fluctuation

RNA RiboNucleic Acid

List of Acronyms

RNN Recurrent Neural Network

RTK Receptor Tyrosine Kinase

SVM Support Vector Machine

TBM Template Based Modelling

TK Tyrosine Kinase

TN True Negative

TP True Positive

vdW Van der Waals

WT Wild Type

Peer Reviewed Publications

Pfeiffenberger, E., Chaleil, R. A. G., Moal, I. H., and Bates, P. A. (2017). A machine learning approach for ranking clusters of docked protein-protein complexes by pairwise cluster comparison. *Proteins: Structure, Function and Bioinformatics*, 85(3):528–543

CHAPTER 1

Introduction

> "Perhaps the most remarkable features of the molecule are its complexity and its lack of symmetry. The arrangement seems to be almost totally lacking in the kind of regularities which one instinctively anticipates, and it is more complicated than has been predicated by any theory of protein structure. Though the detailed principles of construction do not yet emerge, we may hope that they will do so at a later stage of the analysis."
>
> *A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis – Kendrew et al. (1958)*

1.1 Thesis Outline

In this thesis, proteins and their interactions with other proteins are treated as dynamic systems, that are flexible and adapt to different states. With this concept in mind several *in-silico* methods are developed and applied to predict their fold, their binary interactions and their dynamics with atomistic detail.

In Section 1.2 of this introduction chapter, a justification of my research is given by providing examples of possible applications that will help to accelerate scientific discovery and guide drug design. The rest of the chapter introduces proteins as a mechanical system and what constitutes their interactions. This

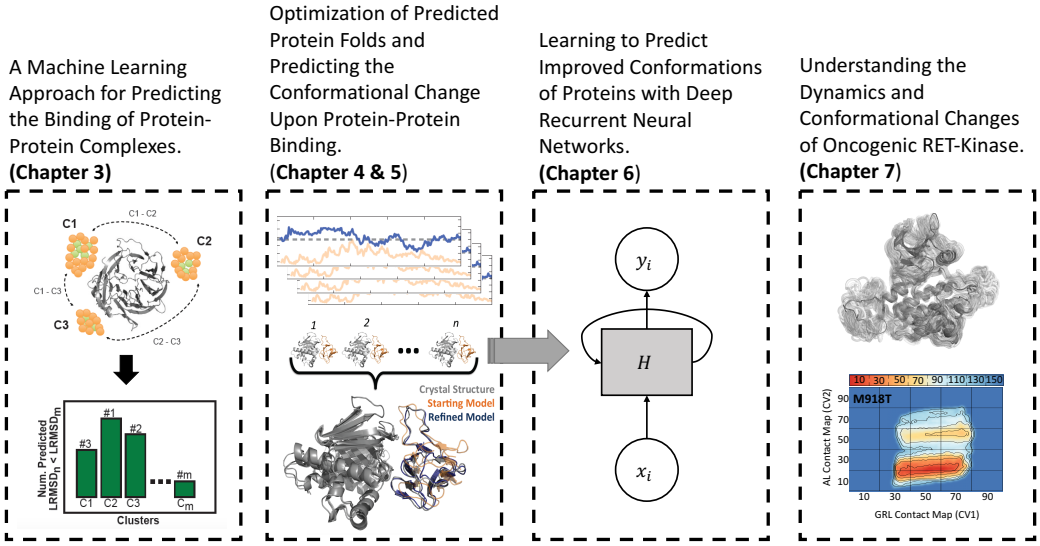

Figure 1.1: Graphical abstract. In this thesis, results are presented from the work on ranking clusters of docked protein-protein complexes by describing local binding sites with distributions from a large number of molecular descriptors and integrating this into a machine learning model (Chapter 3). In the following two chapters a refinement method was developed to improve the model quality of protein monomers (Chapter 4) and predicting the induced conformational change from unbound to bound in protein-protein complexes (Chapter 5). Subsequently, a spatial-temporal model is developed trying to answer the questions can we predict when a protein folds closer to its native state? Given the energy and movement patterns as a function of time (Chapter 6). And finally, a working model of RET-kinase function is proposed from large scale in-silico biophysical experiments. Illuminating how the concerted dynamics of the glycine rich loop and the activation loop defines the auto-phosphorylation trajectory in wildtype and the oncogenic mutant M918T (Chapter 7).

is followed by a discussion into issues and current limitations in structure and interaction prediction (Section 1.3). The concepts behind machine learning and deep learning are introduced and their applications in protein-science are reviewed (Section 1.4). Finally, an overview of the RET-kinase is given (Section 1.5).

In Chapter 3 a novel protein docking ranking method is presented that models the dynamic recognition process of protein-protein binding, where different conformations and binding modes are explored, by quantifying the local energy landscapes with a large number of molecular descriptors and a pair-wise learning strategy to distinguish incorrect from correct binding sites.

Chapter 4 addresses the problem of improving the model quality of predicted

protein folds by refinement. A method was developed that performs refinement by restrained molecular dynamics (MD) simulation and in a latter extended version performs MD sampling in a so called contact map space (CMS) of residue-residue contacts to sample different configurations of the systems more directly. This method was extended to protein-protein complex refinement in Chapter 5, where the CMS is defined between interface residues at the binding site. Furthermore, a new scoring function was formulated that combines statistical potential terms with the reconstructed free energy from CMS-MD sampling. Following this extensive exploration into sampling protein monomers and dimers as a function of time, the following question was addressed: is it possible to predict when a protein folds more closely to its native state? To that end a spatio-temporal model based on a deep recurrent network was defined and trained on data from more than 1.7 million time-points. This model learns to classify segments by looking at patterns of energies and distances in time (Chapter 6).

In the last result Chapter, 7, extensive in-silico biophysical experiments are performed to understand the auto-phosphorylation mechanism of RET-kinase in wild-type and its oncogenic mutant form M918T. The function of RET is interpreted from a dynamic protein-motion perspective where the concerted movement of its glycine rich loop and activation loop are necessary for ordered function.

1.2 A Thesis Justified

Predictive modelling of protein structure, their dynamics and their interactions has tremendous value for scientific discovery. In the following subsections possible applications are presented that will benefit from improvements in these areas.

1.2.1 Slow Accumulation of Experimental Structural Data

The determination of protein structure has hugely contributed to our understanding of molecular function. Experimental methods such as X-ray crystallography and more recently electron microscopy allowed us to study their properties with

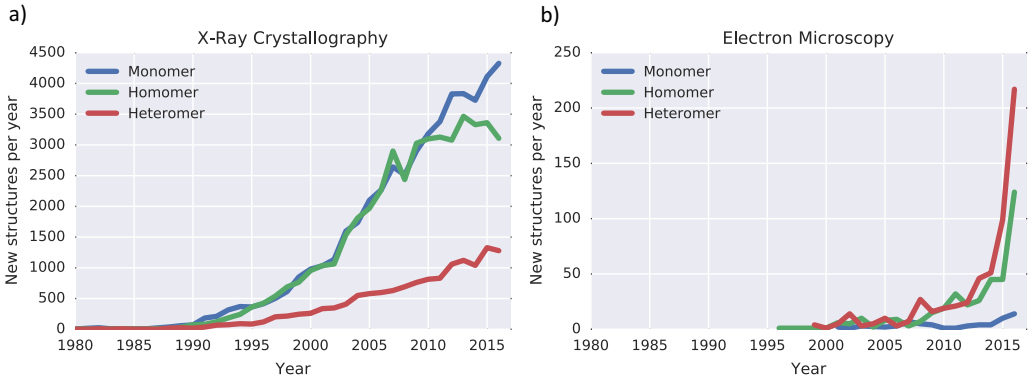

Figure 1.2: Accumulation of experimental structures. Shown are the yearly deposition of new monomers (blue), homomers (green) and heteromers (red) structures for a) X-ray crystallography and b) electron microscopy.

atomistic detail. The major histocompatibility complex (MHC) class I, for example, is required for the immune system response in cells. When disease associated proteins are present in the cell they are cleaved into into small fragments, known as antigens, by the proteolytic machinery and presented to the cell membrane by MHC I (Vigneron et al., 2004). These disease infected cells are recognised by cytotoxic T cells where the interaction between a tri-complex of T-cell receptor, antigen and MHC I mediates the destruction of the cell (Janeway et al., 2001). The structural determination of this tri-complex illuminated the mechanical understanding of the human immune response with T-cells to disease (Krogsgaard et al., 2005).

Another example is the tyrosine specific kinase epidermal growth factor receptor (EGFR). Dysfunction of this transmembrane protein is associated with diseases such as Alzheimer's or different cancer types. In Alzheimer's disease dysfunctional EGFR signalling promotes neurodegeneration associated processes (Repetto et al., 2007) whereas in cancer an over signalling by enhanced kinase activity promotes tumour cell development (Normanno et al., 2006). The resolved structure of EGFR helped to understand these mechanism by providing details of residue-residue interactions that lead to activation of the catalytic domain (Jura et al., 2009) as well as evidence for the function of the extracellular signalling receptor in wild-type and mutant structures (Ferguson, 2008).

However, the experimental determination of the atomic structure of protein

systems is restrained and often less than perfect. Factors that make the determination of protein structure by X-ray crystallography challenging are the need for large quantities of purified protein and successfully growing the crystal for X-ray diffraction. The construction of a functional expression system to produce the quantities required for X-ray crystallography are time-consuming and can take several years. Once this step is accomplished it is still not clear whether the protein will crystallise. Especially challenging are transmembrane proteins where even despite great progress in experimental protocols (Chen et al., 2010; Miller and Long, 2012; Rasmussen et al., 2011) many of them are still left unsolved. Another layer of complexity is added by the transient nature of most heterogeneous protein-protein interactions (Perkins et al., 2010). The instability of their interaction often lead to failed crystal growth, that requires stability of the protein system to be successful. This is reflected in the annual growth rate of newly deposited structures, where the accumulation of new heteromer structures is small (Figure 1.2a). This problem is partially solved by electron microscopy that allows to solve larger heteromer complexes. This is a promising new technology that experienced a rapid growth in deposited structures from 2014 onwards (Figure 1.2b). Initially, the solved structures were of low resolution and lacked the detail required to understand function. However, with continued improvements of the method over the years, resolutions that are close to x-ray crystallography can be obtained for many heteromer complexes (Gao et al., 2016; Su et al., 2017; Škubník et al., 2017). Nevertheless, the structurally solved protein interaction-space by x-ray crystallography and electron microscopy is small in comparison to the known protein interactions, e.g. as annotated in the IntAct protein interaction databank (Kerrien et al., 2012). Thus, it is important to push improvements of computational methods that allow for accurate prediction of protein structure and interaction with atomic detail for a wide range of protein systems.

1.2.2 Rational Drug Design

The case for accurate structure prediction methods is not only given for a mechanistic understanding of protein function but also for targeted drug design for disease associated proteins. The knowledge of high resolution structures of drug targets enables the efficient design of high affinity small molecule compounds that can act as inhibitors in order to disrupt disease associated enzyme activity or protein-protein interaction.

A classical example of successful drug design guided by molecular modelling is the development of AIDS (Acquired Immune Deficiency Syndrome) drugs that act as inhibitors for the protease and the reverse transcriptase of the human immunodeficiency virus (HIV). This retrovirus infects human cells by transcribing its ribonucleic acid (RNA) into deoxyribonucleic acid (DNA) with the reverse transcriptase enzyme. The DNA product is then incorporated into the hosts genome by the integrase (Wilen et al., 2012). The development of current inhibitors was possible due to the structural knowledge from X-ray crystallography. This knowledge allowed for targeted design of protease inhibitors, such as Indinavir, Saquinavir and Ritonavir, that disrupt the viruses life-cycle by blocking the activity which is essential for the virus to mature, reproduce and become infectious (Chen et al., 1994). Likewise, the design of reverse transcriptase inhibitor azidothymidine was enabled by structural knowledge (Ren et al., 1998).

In an *in-silico* drug-discovery study of the human estrogen receptor α, several promising high affinity compounds could be identified (Sivanesan et al., 2005). This study exploited 3 ns MD simulations to model the receptor flexibility with 51 conformational states. A docking study of a 3500 compound library was performed where each compound was docked into each conformational state. The authors argue that the modelled flexibility was necessary to correctly identify some of the compounds.

More recently, interest has increased in the development of protein-protein interaction inhibitors. These have long been seen as undrugable due to shallow binding interfaces and the large surface area of binding sites in

protein-protein interactions, that as a result makes the design of high affinity small molecules hard(Mullard, 2012). In a study by Filippakopoulos et al. (2010) the small molecule inhibitor JQ1 was shown to prevent the interaction between bromodomain-containing protein 4 (BRD4) and acetylated histone proteins. The inhibition of the interaction of this epigenetic reader was *in-vivo* validated by the authors to have potential therapeutic applications for human squamous carcinoma (Miyoshi et al., 2001; French et al., 2003).

1.2.3 Protein Function is Coupled to Dynamics

Proteins are dynamical systems that perform their function and interaction with other proteins by concerted movements of their structural building blocks. One of the most marked shortcomings of X-ray crystallography is that it only captures one static state of the protein with the result that certain functions can't be well understood. This was, for example, the case for the RET-kinase studied in Chapter 7, where the crystal structure of wild-type and oncogenic mutant failed to explain the regulation of the auto-phosphorylation trajectory. The relationship between dynamics and regulation of activity has been shown to matter for a wide range of different kinases.

A study by Foda et al. (2015) on Src kinase function showed how an allosteric network of dynamically linked residues connects ATP- and substrate-binding sites to regulatory sites. In their MD simulations from an active state, spontaneous transitions to an inactive state were observed. This involved concerted movements of several key structural elements in the catalytic domain. From this they could suggest that the phosphoryl transfer during the catalytic cycle causes a switch in the allosteric network with the effect of different stable conformations that have distinct substrate-binding characteristics.

An MD simulation of the cyclin-dependent kinase 5 (CDK5) investigated the transition of its activation loop from an active state to an inactive state (Berteotti et al., 2009). In their simulations the free energy landscape was reconstructed and two new intermediate states were discovered that have not been reported before in

crystal structures. The authors suggest that this information would be valuable for inhibitory drug design in order to trap CDK5 into inactive conformations.

1.3 The Physics of Proteins

1.3.1 The Framework of Protein Conformation

The wide range of different three dimensional protein folds as well as their collective back-bone motions in time can be described in terms of their dihedral rotations around ϕ and ψ. Figure 1.3a shows a schematic illustration of a tri-residue fragment where the ϕ and ψ rotation is defined around N-Cα and Cα-C, respectively. The actual peptide bond from a residue i to $i+1$, connected by atoms C-N, is rigid, thus, the dihedral rotation referred to as ω is not possible due to the double bond like properties of this atom-pair. Although, ϕ and ψ are flexible bonds, they are restrained to certain rotational regions as imposed on them by the formation of secondary structure elements such as α-helices and β-sheets. Figure 1.3b shows these two large clusters in a Ramachandran plot of ϕ-ψ distribution from a set of 500 high-resolution crystal structures.

Side-chain flexibility is described by dihedral rotations around χ where 18 out of the 20 amino-acids (excluding alanine and glycine) can adapt different conformational states known as rotamers. In Figure 1.3d an example is shown for lysine, with possible rotations around $\chi_1, \cdots, \chi_4$ that define the rotations around atom bonds Cα-C1, C1-C2, C2-C3 and C3-C4, respectively.

1.3.2 Proteins as a Mechanical System

From a physics perspective, a protein system can be simplified into a mechanical system where atoms are connected by springs and where physical forces induce the three dimensional structure and collective motion of its atoms. A quantification of

[1] `https://en.wikipedia.org/wiki/Ramachandran_plot`, Last accessed: Oct 23 2017.

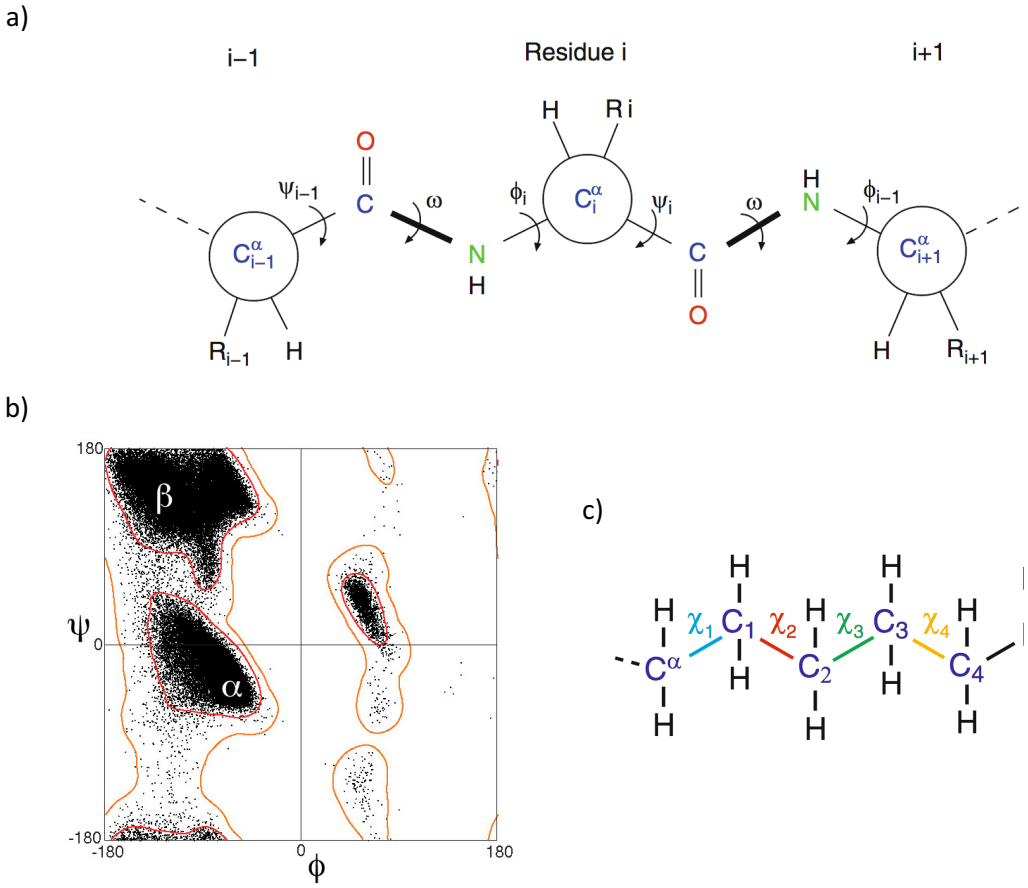

Figure 1.3: Framework of protein conformation. a) Schematic illustration of the backbone of a tri-residue fragment, the side-chain is not shown and only indicated by R_{i-1}, R_i and R_{i+1}. The two possible rotations are marked by ϕ and ψ and the rigid rotation by ω. b) Ramachandran plot of ϕ-ψ distribution in 81,234 residues (proline and glycine excluded) from 500 high-resolution crystal structures (data from Lovell et al. (2003)). The countours marked by β and α show favoured β-sheets and α-helices ϕ-ψ values, respectivley. c) The 4 possible side-chain rotations χ_1, χ_2, χ_3 and χ_4 for residue lysine. Sub-figures a and c reproduced from Schlick (2010); sub-figure b reproduced from Wikipedia[1]. Permission to reproduce Figures a and c has been granted by Springer Nature.

a protein's energy, E, can be obtained by a series of additive terms:

$$E = E_{\text{bond}} + E_{\text{angle}} + E_{\text{torsion}} + E_{\text{improper}} + E_{\text{vdw}} + E_{\text{es}}. \tag{1.1}$$

In this equation the terms E_{bond}, E_{angle}, E_{torsion}, E_{improper}, E_{vdw} and E_{es} describe the energetic contribution of the bond between two atoms, the angle spanned between three atoms, the torsion spanned between four atoms, the improper torsion angles between four atoms, the Van der Waals (vdW) force between atoms and the electrostatics, respectively. This simplification of the real physics of a system is known as molecular mechanics and relies on three principles: *i.* the thermodynamic assumption, *ii.* additivity of the energy terms and *iii.* transferability.

The thermodynamic assumption states that the folded state of a protein is naturally obtained given the assumption that a protein system folds back into a low free energy configuration, known as the native state, when unfolded into a high free energy configuration, known as the denaturated state. Furthermore, this folding process from denaturated to native state must follow a well defined folding-pathway. Cyrus Levinthal, formulated the hypothesis, known as the Levinthal paradox, that a random search of all possible protein conformations would not yield a native folded state since this enumeration would require years for a protein to fold correctly, compared to the observed sub-second to second time-scale of protein folding that occurs in nature (Levinthal, 1968, 1969).

The assumption of additivity states that the different energetic contributions to the total energy of a protein, as shown in Equation 1.1, can be formalized as the sum of their parts. This decomposition of the total energy has also practical implication such as that certain terms can be evaluated less often in a simulation to reduce the overhead of calculations.

The last assumption, transferability, specifies that the correctness of Equation 1.1 applies to all protein systems. This is possible due to the observation that the same chemical subgroups in a large variety of different proteins adapt the same values. For example, the bond length of backbone atoms Cα-N is near identical in all proteins and thus transferability is given. In the following subsections the

functional form and an explanation of all terms of Equation 1.1 is given.

1.3.2.1 Bond Potential

The bond length potential denoted as, E_{bond}, models the bond length variation and its associated penalty energy from a reference value such that

$$E_{\text{bond}} = \sum_{bonds(ij)} \frac{k^{(ij)}}{2} \left(r_{ij} - r_0^{(ij)} \right)^2, \tag{1.2}$$

where r_{ij} is the bond length between atom i and j, $r_0^{(ij)}$ the reference value and $k^{(ij)}$ a constant. The reference value for $r_0^{(ij)}$ and the constant for $k^{(ij)}$ are derived from observed bond lengths in X-ray crystal structures and the measurement of mass and frequency of a particular bond vibration (Schlick, 2010). Equation 1.2 is the harmonic description of the bond potential and is favoured because of its efficient computation. An illustration how the energy changes as a function of bond length r_{ij} is shown in Figure 1.4a.

1.3.2.2 Bond Angle Potential

The bond angle potential, E_{angle} , describes the associated energy increase in bond angle variation between three atoms such that

$$E_{\text{angle}} = \sum_{angles(ijk)} \frac{k^{(ijk)}}{2} \left(\varphi^{(ijk)} - \varphi_0^{(ijk)} \right)^2, \tag{1.3}$$

where $\varphi^{(ijk)}$ is the angle between atoms i, j and k, $\varphi_0^{(ijk)}$ the reference value and $k^{(ijk)}$ a constant. A visualisation of this potential is shown in Figure 1.4b.

1.3.2.3 Bond Torsion Potential

The bond torsion potential, $E_{torsion}$, is a multi-minima potential defined as

$$E_{\text{torsion}} = \sum_{torsions(ijkl)} \frac{k^{(ijkl)}}{2} \left(1 + \cos(n^{ijkl}\tau - \tau_0^{(ijkl)}) \right)^2, \tag{1.4}$$

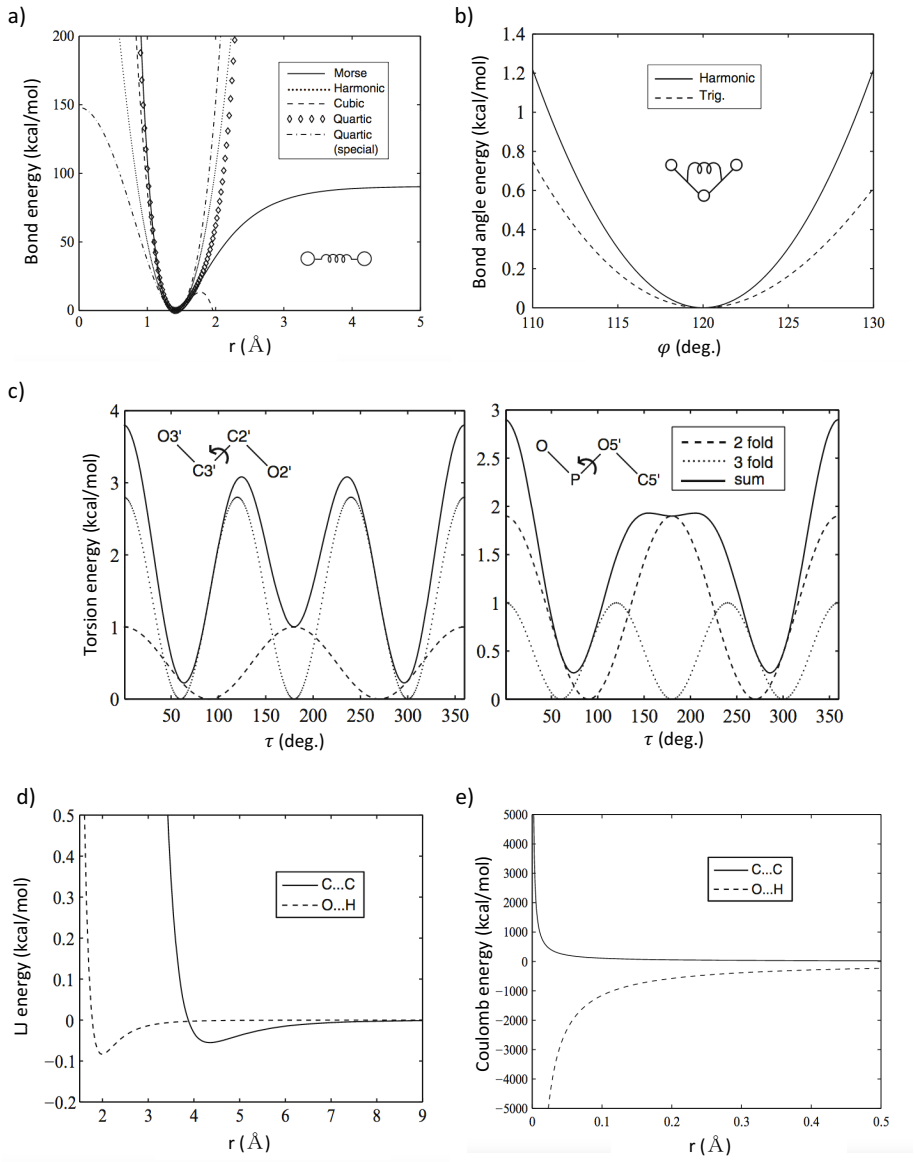

Figure 1.4: Potential energy terms for proteins. a) Bond energy functions Morse, harmonic (visualisation of Equation 1.2), cubic, quartic and quartic special form. b) Bond angle energy for functions of harmonic (visualisation of Equation 1.3) and trigonometric form. c) Torsion energy from the two-fold function, three-fold function and their sum (Visualisation of Equation 1.4). The left hand side shows the torsion energy for nucleic-acid riboses and and the right hand side for rotation around P-O in nucleic-acids. d) Van der Waals energy as described by the Lenard-Jones potential (visualisation of Equation 1.6) for paired interactions of C-C and O-H atoms. e) Electrostatic energy as described by the Coulomb potential (visualisation of Equation 1.7) for interacting atom pairs C-C and O-H. Figure reproduced from Schlick (2010). Permission to reproduce this Figure has been granted by Springer Nature.

where τ is the torsion angle, $\tau_0^{(ijkl)}$ the reference value and $k^{(ijkl)}$ a constant. Additionally, the potential has the non-negative integer parameter n^{ijkl} which describes the periodicity.

1.3.2.4 Bond Improper Torsion Potential

The bond improper torsion potential, E_{improper}, describes the increase in energy associated with the variation of the improper torsions spanned by four atoms such that

$$E_{\text{improper}} = \sum_{impropers(ijkl)} \frac{k^{(ijkl)}}{2} \left(\xi^{(ijkl)} - \xi_0^{(ijkl)}\right)^2 \tag{1.5}$$

where $k_{(ijkl)}$ is the improper dihedral force constant, $\xi^{(ijkl)}$ the improper dihedral angle and $\xi_0^{(ijkl)}$ the reference value. This term ensures the planarity of aromatic rings in amino-acids such as histidine, proline, phenylalanine, and tryptophan.

1.3.2.5 Van der Waals Force

The van der Waals (vdW) force is described as a quickly decaying potential as the distance between two nonbonded atoms increases:

$$E_{\text{vdw}} = E_{\text{LJ}} = \sum_{pairs(ij)} \left(\frac{-A_{(ij)}}{r_{ij}^6} + \frac{B_{(ij)}}{r_{ij}^1 2} \right). \tag{1.6}$$

This functional description is known as the Lennard-Jones potential (Jones, 1924) where the attractive ($-A_{ij}$) and the repulsive ($B_{(ij)}$) part part are dependent on the distance ($r_i j$) and atom type of two atoms. An illustration of this potential is shown in Figure 1.4d.

1.3.2.6 Electrostatic Contribution

The electrostatic contribution to the potential energy, E_{es}, is described by the Coulomb potential such that

$$E_{\text{es}} = E_{\text{coul}} = \sum_{\text{pairs}(ij)} \frac{1}{4\pi} \frac{q_i q_j}{\varepsilon r_{ij}}, \tag{1.7}$$

where r_{ij} is the distance between two atoms, ε the dielectric constant and k_{coul} a conversion factor to obtain energies in kcal/mol. The effect charge for atom i and j is expressed by q_i and q_j, respectively. Where two positive or negative charges result in a repulsion and in the case of two opposite charged atoms into an attraction. An example of this is shown in Figure 1.4e.

1.3.3 From Sequence to Structure: Issues and Current Limitations in Structure Predictions

Since the first protein structure was solved, considerations began whether the principles of protein folding could be learned to derive the three dimensional structure from its amino-acid sequence alone. Despite the ever increasing number of new protein structures that have been experimentally solved, no computer algorithm could be compiled that allows for accurate *ab-inito* folding of all proteins.

The most reliable structure prediction methods draw from the observation that proteins with a similar sequence also possess a similar structure. Work by Rost (1999) analysed more than 1 million pair-wise sequence alignments to identify at what level of sequence dissimilarity the structural similarity is no longer given. He concluded that with a sequence similarity of 30 percent or more, 90 percent of all pairs had a homologous structure. Below a threshold of 25 percent, the relative number of homologous pairs dropped to 10 percent. Template based methods (Meier and Söding, 2015; Biasini et al., 2014) exploit this observation by searching for sequence homologous proteins with solved structures. This approach becomes more successful as more structures are experimentally resolved. Currently, 124782

structures are available in the Protein data bank (PDB; Berman et al. (2000)). However, many of these structures are similar, with only 1375 reported unique folds[2].

A large number of protein families remain structurally un-described, where for 4600 Pfam families no single crystal structure is available (Söding, 2017). This "dark" protein space poses a serious problem for structure prediction with template based methods. The gap is filled by *ab-inito* methods that predict the structure from assembling small structural fragments (Simons et al., 1997). Here, the target protein is split into small overlapping sequence fragments and queried against a structural fragment database to find matches that are assembled together into one structure. These methods have limited success at correctly predicting the fold and are usually only successful for smaller proteins with 100 residues or less. A problem associated with fragment based predictions are the scoring functions that rank the solutions, which often fail to identify the right fold (Moult et al., 2016).

An alternative method for *ab-initio* structure predictions are MD folding simulations, where starting from a denaturated state the funnel is descended to the native state (see Figure 1.5a). It has been shown that such folding simulations are successful for small proteins. Work by Lindorff-Larsen et al. (2011) showed that 12 proteins could be successfully folded into their native state. However, folding simulations of larger proteins were not successful at producing the native state. A crossover between folding simulations and structure prediction methods are refinement methods. Here, the simulation starts from an initial pre-folded model and conformational sampling with MD is performed to decent to the bottom of the folding funnel (see Figure 1.5b). Work by Raval et al. (2012) showed that long unrestrained simulations of 100 μs do not lead to improvements in model quality and rather drift away. A method by Mirjalili et al. (2014) that made use of short restrained MD simulations performed much better and was the most successful refinement method in CASP11 (Modi and Dunbrack, 2016). The main limitations of folding or refinement by simulation are force field inaccuracies, high computational

[2] `http://www.rcsb.org/pdb/statistics/holdings.do`, Last accessed: Oct 23 2017

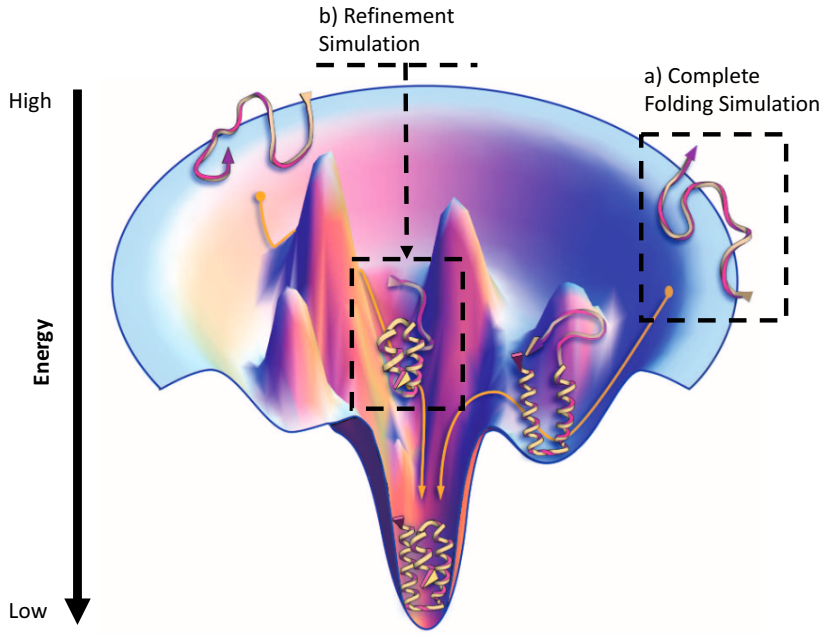

Figure 1.5: Protein folding funnel. Illustration of the folding funnel where many high energy conformations are possible but only a few low energy conformations. a) Illustration of complete folding simulations from a completely unfolded state. b) Illustration of a refinement simulation that starts from a pre-folded state and tries to decent the latter part of the funnel only in order to reach the native state. Figure adapted from Dill and MacCallum (2012). Permission to reproduce this Figure has been granted by AAAS.

cost, and current limitations at finding the snapshots in the trajectory that resemble the native structure best (Feig and Mirjalili, 2016).

More recently, residue-residue contact prediction from co-variance analysis of large sequence alignments (Jones et al., 2012) improved dramatically in precision (Jones et al., 2015; Wang et al., 2017a). The predicted residue-residue contacts are used as restraints in model building which allowed for improvements in model accuracy for large *ab-initio* targets. One short-coming of these methods is that large sequence alignments of several hundred to thousand sequences are necessary, essentially limiting their applicability to only 400 Pfam families for which no crystal structure is known (Söding, 2017). Work by Ovchinnikov et al. (2017) showed that this shortcoming could be overcome by including metagenome sequence data. This approach increased the number of Pfam families with large enough sequence alignments to 1300.

1.3.4 Protein-Protein Interactions

The previous sections discussed the underlying principles that allow for protein flexibility and their physical forces as approximated by molecular mechanics that explain the folded states that are observed in crystal structures and can describe the collective motions. However, most proteins do not live in isolation, they are rather involved in interactions with other proteins important for cell function. These interactions are a complex process that involves in almost all cases conformational transition from an unbound conformational state to a bound conformational state. Such conformational changes range from side-chain rotations to large back-bone movements. The binding process is illustrated in Figure 1.6 where in sub-figure a an unbound ligand samples in the absence of the receptor different conformational states, this process is denoted as conformational selection. When the ligand and receptor are close in space, see sub-figure b, the so called encounter complex is initiated that induces further conformational changes, referred to as induced fit. This mechanism is important for the recognition of the correct binding site where, if correctly bound at the right side, stabilising contacts are induced between the receptor and ligand which lead to a stable complex (sub-figure c). In the other case, if the encounter complex is at the incorrect binding site, the interaction does not induce a stabilizing conformation, hence leading to disassociation.

1.3.4.1 A Kinetics's Description of Protein-Protein Interaction

The complex formation C between two proteins, referred to as receptor R and ligand L can be described from a kinetics's perspective:

$$R + L \rightleftharpoons C. \tag{1.8}$$

From this, two processes can occur. The first case, describes the decomposition of complex C into its monomers R and L and the second case, the composition of R and L into its complex C. The rate of these processes is dependent on the relative concentrations of R, L and C in solution which are denoted as $[R]$, $[L]$

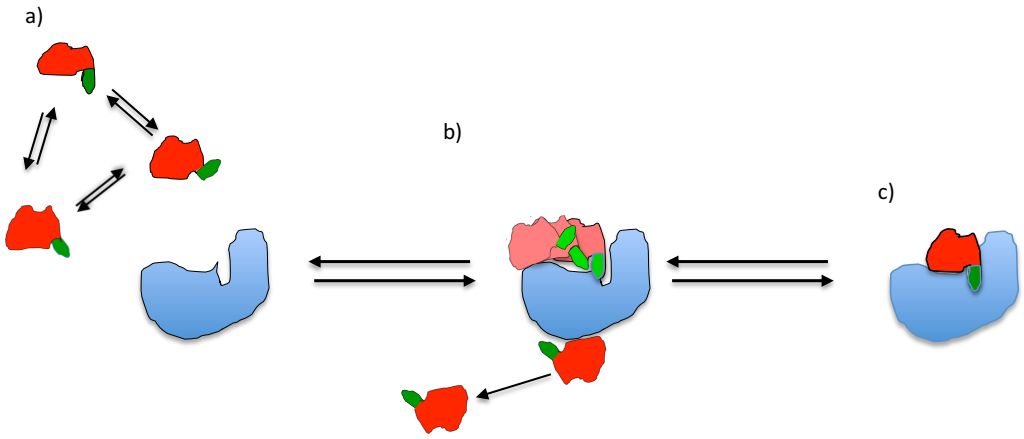

Figure 1.6: Schematic illustration of the process of protein-protein recognition and binding. a) The ligand depicted in red with a flexible hinge shown in green samples different conformational states in the absence of the receptor seen in blue. This mechanism is known as conformational selection. b) The interaction between the receptor and ligand during the recognition process induce further conformational changes. The interaction at the correct binding side shown shown on the upper half allows to the ligand to transition into a state which leads to stable binding. Whereas the interaction on the incorrect binding side, shown at the bottom half, does not lead to a stable association of receptor-ligand and leads to unbinding. This process is known as induced fit. c) Final bound state between receptor and ligand.

and $[C]$ respectively. Modelling of these events can be described in terms of their association and disassociation rate such that:

$$\text{association rate} = k_{\text{on}}[R][L] \tag{1.9}$$

and

$$\text{disassociation rate} = k_{\text{off}}[C]. \tag{1.10}$$

A system is in equilibrium when

$$k_{\text{off}}[C] = k_{\text{on}}[R][L]. \tag{1.11}$$

Consequently, constants k_{off} and k_{on} are the same and the concentrations of $[R]$, $[L]$ and $[C]$ remain unchanged. The stability of a complex is given by its affinity that directly relates to the linear combination of $[R]$ and $[L]$ to its ratio $[C]$:

$$\frac{[R][L]}{[C]} = \frac{k_{\text{on}}}{k_{\text{off}}} = K_D. \tag{1.12}$$

This affinity K_D can be explained in terms of the energetic contribution of the non-bonding terms E_{vdw} (Equation 1.6) and E_{es} (Equation 1.7) which mainly contribute to the stability of a complex.

1.3.5 From Monomers to Dimers: Issues and Current Limitations in Protein Complex Prediction

In-silico prediction of protein-protein interactions is the only feasible choice, given the slow accumulation of experimental structures of protein heterodimers and the huge number of possible protein-protein pair combinations. The fact that most protein-protein interactions occur together with conformational transition poses a problem that often leads to falsely identified binding sites. This is the case when the scoring function used is not able to correctly quantify the binding energy. Figure 1.7a illustrates this schematically. Here, the estimated binding energy landscape by a scoring function is not able to correctly describe the binding energy landscapes due to a conformation that is different to the bound state. An example is given in Figure 1.7b which shows the scoring of docked solutions for the scoring function DCOMPLEX (Liu et al., 2004). The cartoon shows the binding mode observed in the experimental crystal structure. In this case the lowest energy solutions are predicted to be at an incorrect binding site. Work by Kuroda and Gray (2016) quantified the extent of how much conformational deviations to the bound state influences the docking success rate, the results are shown in Figure 1.7d for three scoring functions. The authors define the docking success rate by the presence of an energetic binding funnel (details can be found in Chaudhury et al. (2011)). All three functions experience a dramatic drop in docking success rate with slightest

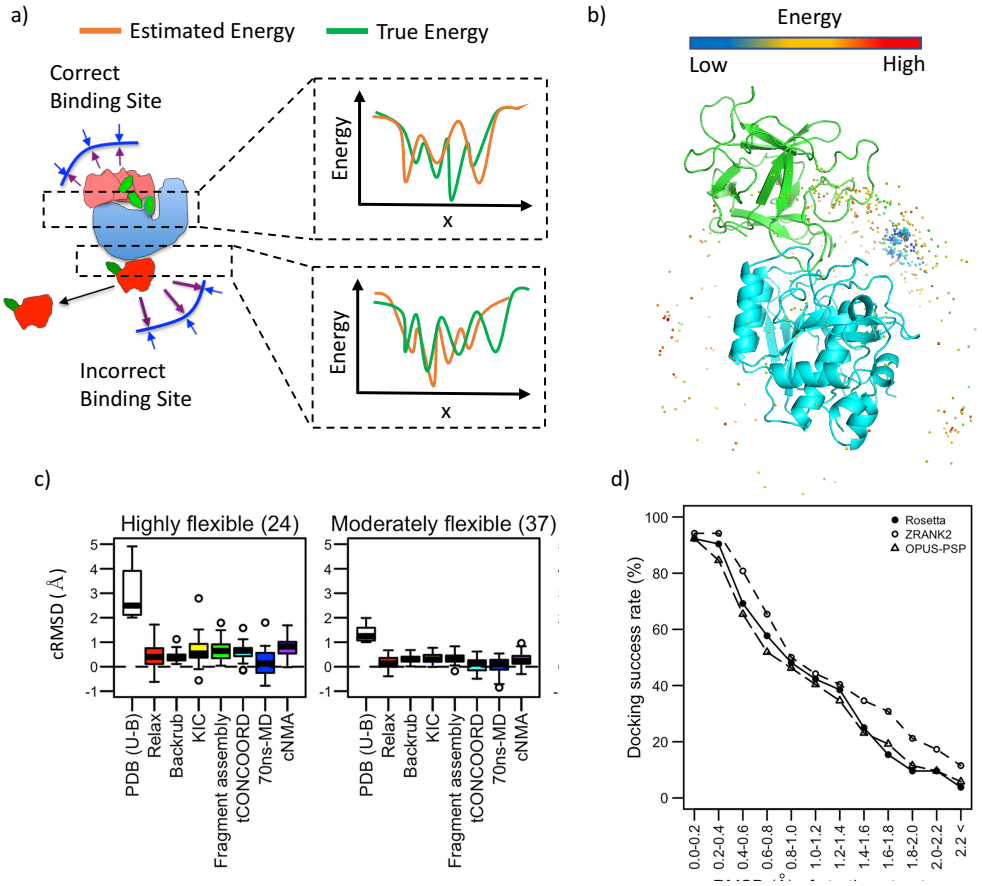

Figure 1.7: Difficulties of predicting protein-protein interactions. a) Difficulties in predicting the correct protein-protein binding site arise mainly from unbound to bound conformational transitions involved in the binding process. Classical energy functions, whether it is a physical or statistical potential, have problems estimating (orange energy landscape) the true binding energy (green energy landscape) if the bound conformational state is not known. Thus, often resulting in lower estimated energies for incorrect binding sites than the correct binding site. b) Example of the energy of docked solutions for CAPRI target T32. The solutions are shown as spheres, where each sphere represents the centre of mass of the ligand position. The color visualizes the DCOMPLEX energy ranging from blue (low energy) to red (high energy). The cartoon rendering in blue (receptor) and green (ligand) is the observed bound complex in the crystal structure. c) Quantification of the ability to sample the unbound to bound transition (for high) in protein binding for 7 different methods, split into two groups, highly flexible complexes with large transitions (left) and moderately flexible complexes with less conformational change (right). The observed conformational RMSD (cRMSD) in the PDB structures between unbound (U) and bound (B) is shown in white. d) Docking success rate as a function of cRMSD for three scoring functions ZRANK2 (Pierce and Weng, 2008), Rosetta (Lyskov and Gray, 2008), OPUS-PSP (Lu et al., 2008). Docking success rate is defined by the presence of an energetic binding funnel, details can be found in Chaudhury et al. (2011). Figures c and d reproduced from Kuroda and Gray (2016). Permission to reproduce Figures c and d has been granted by Elsevier.

conformational differences. For example, a deviation of 1 Å in conformational RMSD (cRMSD) leads to a decrease in docking success rate to about 50 percent. A study by Kuroda and Gray (2016) explored how good current sampling methods are at sampling the transition from unbound to bound state. The results show that all 7 tested methods are not able to fully sample the transition for highly flexible protein-protein interfaces as well as for moderately flexible interfaces (Figure 1.7c).

1.3.6 Quantification of Protein Complexes

The quantification of protein-protein interactions has been accomplished by a large number of different molecular descriptors. The following text provides an overview of the different categories.

Atomic contact and distance potential: These potentials are usually derived from a set of observed inter atom-atom contacts at the interface of a protein complex. The potentials are described by the inverse of Boltzman's law:

$$u(\alpha,\beta,R) = -kT \ln \frac{p_{\text{obs}}(\alpha,\beta,R)}{p_{\text{exp}}(\alpha,\beta,R)}. \tag{1.13}$$

The variables k and T are the Boltzman constant and the temperature, respectively. Probability functions p_{obs} and p_{exp} express the observed and expected probability of two atom types α and β in their distance bin R. Function p_{obs} is derived from observed atom-atom contacts from a set of crystal structures and p_{exp} is the so called reference state for the background probability of such atom-atom pairs. The distinguishing property between potentials in this category is the formalization of the reference state p_{exp}.

Residue contact and distance potential: Similar to atomic contact and distance potentials but instead of a full atomic description of all heavy atoms in the system only a coarse grained definition is used, where for each residue pair only the $C\alpha$ or $C\beta$ distances are considered.

Composite scoring functions: These functions are made up of weighted additive

terms, for example:

$$E_{\text{tot}} = \alpha E_{\text{LJ}} + \beta E_{\text{coul}} + \gamma E_{\text{desolv}}. \tag{1.14}$$

In this example, the total energy, E_{tot}, is composed of terms describing the van der Waals forces via a Lennard-Jones potential (E_{LJ}), the electrostatic charge (E_{coul}), and the desolvation energy after complex formation (E_{desolv}). The coefficients α, β and γ are the associated weights for each term and express the magnitude of their contribution to the total energy. Usually, these weights are optimised based on a set of decoy structures where the objective function is defined as a minimisation of the ranking error. Where the ranking is usually given by the RMSD between a decoy model and its reference crystal structure. An example of such a function is ZRANK (Pierce and Weng, 2007).

Solvation energy functions: Quantifies the energy associated with the displacement of water at the protein-interface after complex formation with another binding partner. The estimation of this energy can be done by probing the solvent accessibility of residue atoms before and after binding.

Hydrogen bonding: Quantification of the stabilising contribution of inter-molecular hydrogen bonds. Such a bond is formed between a donor atom, i.e. a hydrogen atom bound to a negatively charged atom such as oxygen, and an acceptor with a lone electron. These functions consider the distance and angle between the acceptor and donor to quantify their strength.

Van der Waals and electrostatic: This category contains different functional descriptions for the van der Waals force, which describes the attractive and repulsive non-bonding interaction between two atoms as a function of their distance. An example of such a function would be the Lennard-Jones potential (see Section 1.3.2). The electrostatic energy is defined as the attractive or repulsive force between two differently charged atoms or two equally charged atoms, respectively and is defined, for example, by the Coloumb potential (see Section 1.3.2).

1.4 Machine Learning

The practice to formulate hypothesises and predictions from observational data is as old as science itself, and the process of finding knowledge in data by looking at their patterns has a long history. For example, the astronomer Johannes Kepler studied in the 16th century extensively the patterns of planetary motion from observational data that led to the formulation of Kepler's law that allowed to predict the position of planets (Murray and Dermott, 1999). The automation of this pattern recognition process is what is called machine learning (ML) and is essentially the science of engineering computer algorithms that learn. The automation of this learning process that maps observational data to a predicted outcome has become an important research field that has, and will impact many areas in science and society. The dream of creating so called artificial intelligence (AI) with machine learning algorithms started with the beginning of computers (Turing, 1950) and were picked up in popular culture (Asimov, 1950). However, the following sub-sections give a more immediate introduction to this topic by describing the concepts of statistical and deep learning. A technical description of particular machine learning algorithms that have been used in this work is given in Chapter 2: "Materials and Methods".

1.4.1 The Fundamentals

Essentially, in machine learning a mapping is learned such that

$$f(X) = Y, \tag{1.15}$$

where X is the input, Y is the desired output and the mapping function $f(X)$ is unknown and has to be derived from the patterns in the data. The constitute variables, $X_1, \cdots, X_n$ of the data X are known as the features that describe an entity. For example, an entity could be a protein system which is described by different potential energy terms such that $X_1 = E_{\text{bond}}, \cdots, X_n = E_{\text{vdw}}$.

The different algorithms in machine learning can be divided into two categories that are supervised and unsupervised learning (see Figure 1.8). In supervised

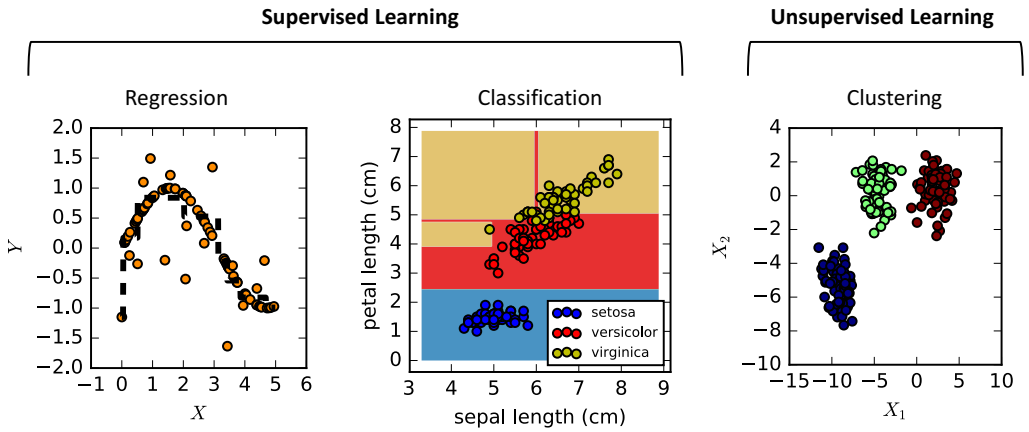

Figure 1.8: Examples for supervised and unsupervised learning. In supervised learning, a machine learning model learns the mapping from input X to output Y from labelled training data. This is divided into learning a regression or classification task. In regression learning the output Y is a continuous variable, the example shown in the left hand panel visualizes how a regression decision tree (predictions shown as a black dotted line) was trained to fit the output function (orange dots). In classification learning the output variable Y is categorical. The example shown in the centre panel shows the classification of three different plant types (setosa, versicolor and virginica) based on their sepal length and petal length from the Iris data-set (Fisher, 1936). The decision surface learned by a decision tree is shown as the coloured background. In unsupervised learning the output Y is not learned from labelled data but has to be directly derived from the shape of the data. The example in the right hand panel shows k-means clustering to group data-points into $k = 3$ classes.

learning, the training data is labelled, i.e. also referred to as the ground truth where the relationship $X \rightarrow Y$ is known in order to learn the mapping function $f(X)$ for new unseen data. Furthermore, it is distinguished between regression problems and classification problems. For a regression problem the output variable Y is continuous, an example of a regression machine learning task would be the prediction of binding affinities K_d for protein-protein complexes. In classification problems the variable Y is categorical. An example of a classification problem is shown in the centre panel of Figure 1.8. In this example a decision tree classifier is trained to predict whether a plant is setosa, versicolor or virginica based on their input features sepal and petal length. During training from labelled example data a decision surface is learned that allows the classification of new unseen value pairs of sepal and petal length into one of the three categories. In the case of unsupervised

Figure 1.9: Machine learning overview. a) Flowchart showing how rule-based systems, classic machine learning, representation learning and deep learning learn from data. Gray boxes represent automatic parts learned from the input. b) Example of how a deep learning model learns more abstract features from image pixels. Figure reproduced from Goodfellow et al. (2016) and Zeiler and Fergus (2014). Permission to reproduce this Figure has been granted by MIT Press.

learning the training data is unlabelled where the relationship $X \rightarrow Y$ is not given. The mapping function $f(X)$ has to learn from the shape of the data alone. For example, Figure 1.8 right hand panel shows the k-means clustering methods where the assignment of classes for data-points depend on the distance to their neighbours.

1.4.2 Deep Learning

In classical machine learning the input features are hand-designed and require expert domain knowledge to train an accurate model. This can be a problem for certain tasks such as object recognition in images where the engineering of discriminative features is challenging. Finding the right representation of the input data manually such that a machine learning algorithm is successful for a given task is time consuming and often not even possible. A way to address this is representation learning that explores the right feature representation automatically to increase the predictive performance. An example is an autoencoder, that consists of a encoder function which converts the input into a new representation. During the training of autoencoders the encoder function is learned such that most of the information in the data is preserved as well as transformed into a new representation that has convenient properties such as linear separability of two classes. In deep learning representation learning goes one step further where from simpler representations higher order, i.e. more abstract, representations are learned (see Figure 1.9a for a comparison). Coming back to the object recognition problem from image data, a deep learning model would learn from input pixels through a series of hidden layers the identifying representations. In the example shown in Figure 1.9b the first hidden layer learns the concept of edges by recognising the brightness of neighbouring pixels. The second layer learns from the edges of the first layer the concept of corners and contours by combining them together which in turn allows the third layer to detect object parts. This final high level representation is predictive enough to reliably identify different objects.

The underlying architecture of deep learning models are deep artificial neural networks (ANN). In the simplest case, known as a perceptron, an ANN consists only

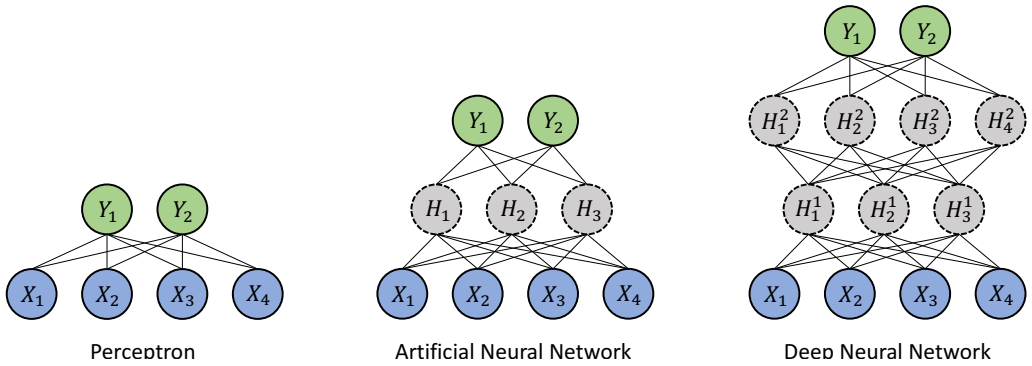

Figure 1.10: High level view of neural networks. Left hand panel shows a perceptron with input nodes directly connecting to output nodes. The centre panel shows an artificial neural network that has an intermediate hidden layer and the right hand panel visualizes a deep neural network with several hidden layer. Nodes coloured in blue are the input, gray the hidden layer and green the output.

of an input layer and an output layer that are fully connected by weighted activation functions. In ANNs with an additional layer, the input is mediated through an intermediate hidden layer to the output. Finally, in a deep neural network multiple hidden layers are stacked. A schematic representation of this concept is shown in Figure 1.10. This simple connection of inputs through a cascade of hidden layers allows, in theory, the learning of any mapping function $f(X)$ (Lin and Tegmark, 2016).

The success of deep learning in a wide range of domains is manifold. An important factor in deep learning is the training-data size. The increased collection of data for many areas in science and society has generated large labelled training datasets that can be leveraged for deep learning (see Figure 1.11, top plot). A rule of thumb is that a convolutional DNN can learn a task such as object recognition with reasonable precision from a set of 5000 images per category and reach human level performance with 10 million images (Goodfellow et al., 2016). Though, this number can vary, and is dependent on the complexity of the problem. However, continued research efforts in areas such as transfer-learning (Raina et al., 2006), generative adversarial networks (Goodfellow et al., 2014) and reinforcement learning (Sutton and Barto, 1998) try to reduce this requirement. Another reason is the increase in computational power that allows to train larger networks on more data. Figure 1.11 bottom plot shows the steady increase in complexity over the

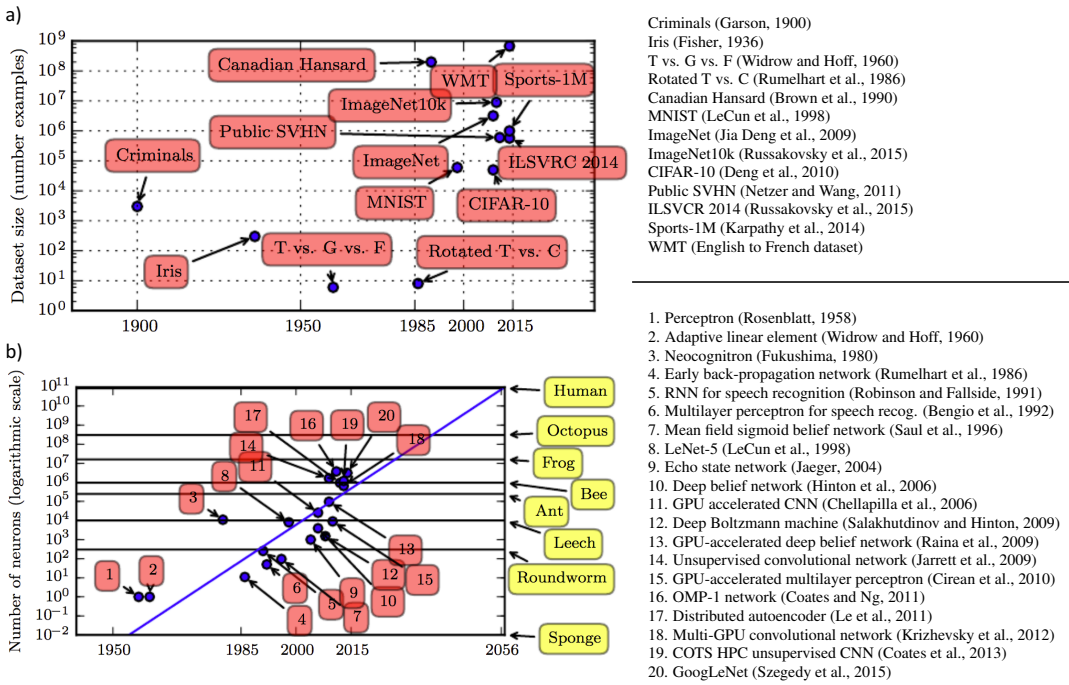

Figure 1.11: Growth of data-set size and neural network complexity. a) The data-set size is constantly increasing from a few hundred to several millions examples of labelled training data. References for the data-sets are shown on the upper right hand side. b) Increase of network complexity from the late 1950th to 2015. References for the different networks are shown on the lower right-hand side. Figure reproduced from Goodfellow et al. (2016). Permission to reproduce this Figure has been granted by MIT Press.

years. The biggest networks, such as GoogLeNet (Szegedy et al., 2015), have more than 1 million neurons and are comparable in complexity to the brain of a bee (Goodfellow et al., 2016).

1.4.3 Applications in Protein Science

In recent years, machine learning and especially deep learning has enabled many breakthroughs ranging from recognising cats in videos (Le et al., 2011) to playing Go better than any human (Silver et al., 2016, 2017) to self driving cars (Bojarski et al., 2016). The current applications of machine learning are so manifold and wide-spread that a comprehensive review is impossible within the scope of this thesis. In the following subsections key advances in different sub-fields of protein-science are presented and discussed.

1.4.3.1 Residue-Residue Contact Map Prediction

The correct prediction of long-range residue-residue contacts has tremendous value for ab-inito structure prediction for cases where no close homologue with known structure is available to apply homology modelling (Kim et al., 2014). Methods based on coevolution analysis (Marks et al., 2011; Jones et al., 2012; Seemayer et al., 2014) have shown success but suffer from the need for large numbers of sequence homologues (de Juan et al., 2013). The method MetaPSICOV (Jones et al., 2015) combined co-evolution patterns together with other protein features into a neural network that has led to a higher precision of correctly predicted contacts in CASP11 (Kinch et al., 2016). More recently, deep convolutional neural networks have enabled a marked improvement in precision where the prediction of contacts is treated as a pixel-level image classification problem that allows to learn complex patterns of co-occurence (Wang et al., 2017a). In CASP12 this method achieved a precision of 0.402 compared to 0.272 and 0.129 for MetaPSICOV (shallow neural network) and CCMpred (standard co-evolution by Seemayer et al. (2014)), respectively (Wang et al., 2017b).

1.4.3.2 Binding Affinity Prediction

Predictors of binding affinity attempt to provide a reliable estimate for quantities related to the stability of protein complexes, for example metrics such as k_{off}, k_{on}, k_d and ΔG (see Section 1.3.4.1). These predictors leverage supervised machine learning algorithms to make inferences based on features derived from structural protein-protein models. That includes features derived from protein-protein docking scoring functions (Moal et al., 2011); residue contacts at the interface and non-interface side (Vangone and Bonvin, 2015); and changes upon binding in enthalpic and entropic energy (Marillet et al., 2016). The current state of the art methods have still their limitations. For example, methods produce reliable estimates for protein-protein interactions with no or little conformational change upon binding, however, as soon as transitions from unbound to bound conformations are observed the accuracy drastically drops (Agius et al., 2013).

Furthermore, currently some methods have issues with good training performance but test performance is lacking (Moal et al., 2011) and poor predictive power for antigen-antibody complexes (Vreven et al., 2012).

1.4.3.3 Scoring in Protein-Protein Docking

Reliable ranking of thousands of protein-protein docking solutions is challenging for classical potential and knowledge based scoring functions (see Section 2.2.2 for a definition) due to the intrinsic heterogeneity and flexibility of their interactions as discussed in Section 1.3.5. Several methods have tried to address this problem from a machine learning perspective. The PROCOS method (Fink et al., 2011), for example, addresses this problem from a probabilistic viewpoint. In their method a support vector machine (SVM) is trained from probability distributions of different energy potentials for near-native/native and incorrect protein-protein complexes to derive a ranking. In SVMs a hyperplane is introduced that seperates the data-points of different categories into two areas. During training, the placement of the hyperplane is chosen such that the gap of the hyperplane to the data-points of each category is as wide as possible (Bishop, 2006). A method developed by Neveu et al. (2016) derives a so called data-driven potential energy function where the coefficients of this function are learned by a SVM. In the IRaPPA method (Moal et al., 2017), ranking solutions is defined as an information retrieval problem, where methods traditionally used in internet search ranking and electoral voting are combined to correctly identify near-native solutions. To that end a large number of molecular descriptors are used as features to train an ensemble of n ranking SVMs (Joachims, 2002) that produce n rankings of a set of solutions where the final ranking is consensus ranking of these.

1.4.3.4 Protein Monomer Quality Assessment

A challenging problem in the blind prediction of protein structure is the assessment of how good the predicted model is. This is especially relevant for *ab initio* structure prediction where no structurally close homologue is available for comparative

analysis. A typical measure for model accuracy is the GDTTS score (see Section 2.6.1 for a definition) that reaches the best value of 100 if the model is in perfect agreement with the crystal structure. In CASP12, several machine learning models were especially successful at this task (Kryshtafovych et al., 2017). The two most successful methods were based on support vector machines (SVM), namely SVMQA (Manavalan and Lee, 2017) and PROQ3 (Uziela et al., 2016). These two methods are similar with respect to the input features they use. Both of them exploit features describing the potential energy, predicted secondary structure and the solvent accessibility of residues.

1.4.3.5 Protein Function Annotation

The fast expanding number of protein sequence data makes the manual annotation of protein function challenging. Function annotation based on extensive experimental validation is desired but not feasible due to the slow rate and associated cost of new function annotation. Thus, the development of reliable predictors is of importance to accelerate scientific discovery (CAFA, 2016). Many of the most successful function prediction methods are based on machine learning and integrate different data-sources. For example, a method developed by Cozzetto et al. (2013) exploits information from diverse data sources such as sequence features, high throughput data from micro-arrays, or tri-gram sequence pattern mining to train a series of support vector machines and neural networks. Another such method is MS-KNN (Lan et al., 2013), that integrates data from sequence similarity scores, protein-protein interactions and gene expression to train a k-nearest neighbour classifier.

1.5 Understanding RET-Kinase

1.5.1 Structure and Function

The RET receptor tyrosine kinase (RTK) consists of a large extracellular domain, an transmembrane region and a intracellular kinase domain (Takahashi et al., 1988)

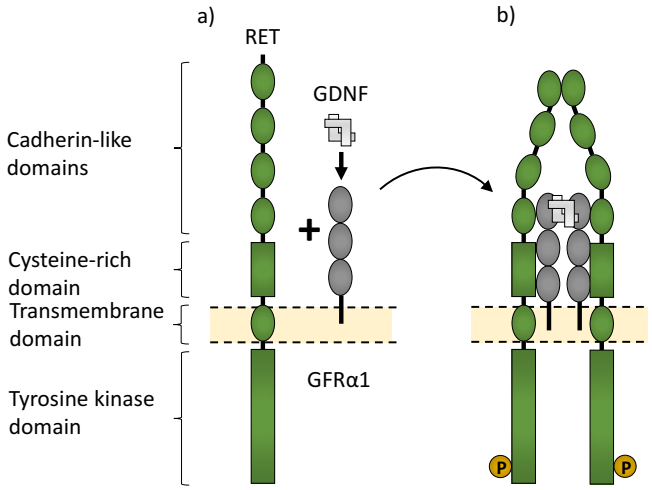

Figure 1.12: RET receptor tyrosine kinase. a) RET is bound to the cell membrane with a long extracellular tail and an intracellular kinase domain. It interacticts with glial cell line-derived neurotrophic factors (GDNFs) such as GDNF. b) RET does not directly interact with GDNFs but via the GFRα co-receptor, upon this the hetero-complex is relocated into lipid raft membrane subdomains and dimerizes. This leads to activiation and auto phosphorylation.

(Figure 1.12a). The extracellular domain is made up of four cadherin-like domains and a cysteine-rich region. These elements are important for dimer stabilisation and co-factor binding (Anders et al., 2001; Kjaer et al., 2010; Amoresano et al., 2005; Wang, 2013). The binding of the glial cell line-derived neurotrophic factor (GDNF) family ligands (GFLs) facilitates signalling in RET. This binding does not happen directly to RET but via a GDNF family receptor-α (GFRα) that are attached to the cell surface by glycosylphosphatidylinositol linkage (Arighi et al., 2005). For example, the dimer complex GDNF and GFRα1 binds to RET and causes a translocation to cholesterol-rich membrane subdomains, also known as lipid rafts, upon which they are activated (Tansey et al., 2000) (see Figure 1.12b). Cross interaction with other GFLs and GFRα receptors provide selectivity for RET activation in different cell types. Furthermore, a variety of other RET interactions, such as other RTKs, adhesion molecules and other cell surface proteins can lead to stabilisation of the signalling complex and enhancement of the catalytic activity (Schalm et al., 2010; Cockburn et al., 2010; Bonanomi et al., 2012; Tufro et al., 2007; Popsueva et al., 2003; Esposito et al., 2008).

The GDNF-GFRα1-RET tri-complex and its subsequent dimerization and autophosphorylation in the kinase domain leads to the recruitment of other adaptor and signalling proteins that trigger downstream pathways. For example, adaptor binding causes the activation of RAS-MAPK and PI3K-AKT signalling pathways or the binding of ubiquitin ligases that downregulates RET function (Hayashi et al., 2000; Besset et al., 2000; De Vita et al., 2000; Coulpier et al., 2002; Segouffin-Cariou and Billaud, 2000).

1.5.2 Biological Importance

RET is most abundant in early embryogenesis and has relatively low levels in adult tissues (Pachnis et al., 1993; Tsuzuki et al., 1995). It is involved in metanephric kidney development where it is required for induction growth, branching and morphogenesis of the ureteric bud (Chi et al., 2009; Davis et al., 2014). Another important function of RET is its involvement in the development and proliferation of the enteric nervous system where it is essential for migration and targeting of neuroblasts to this system (Schuchardt et al., 1994; Durbec et al., 1996). Furthermore, RET is a guidance receptor for axon growth and targeting as well as a neuronal survival receptor in the adult brain and peripheral neurons (Pierchala et al., 2006; Paratcha and Ledda, 2008).

RET is expressed in a wide variety of tissues. A low level RET expression was found during early development of haematopoietic progenitors and an increases in RET expression during myelomonocytic differentiation (Gattei et al., 1997). It could also be shown that RET is expressed in stromal cells in the bone marrow environment (Mulligan, 2014). Furthermore, immune cells such as B cells, T cells, monocytes and natural killer cells are expressing RET (Vargas-Leal et al., 2005; Rusmini et al., 2013).

1.5.3 Cancer Association

RET-associated heritable and sporadic tumours are the result of gain of function mutations that cause an increased expression or activation. The cancer-type multiple

endocrine neoplasia type 2 (MEN2) is associated with germline mutations in RET (Donis-keller et al., 1993; Mulligan et al., 1993). This type of cancer is characterized by early onset medullary thyroid carcinoma (MTC). There are three subtypes known which are Familial MTC (FMTC), MEN2A and MEN2B. FMTC is the mildest form of these subtypes and occurs at a later age of disease onset. MEN2A develops in 50 percent of the cases adrenal tumour pheochromocytoma, in one third of these cases parathyroid hyperplasia or adenoma and/or skin condition lichen planus amyloidosis in addition to MTC (Verga et al., 2003). The subtype MEN2B is the most severe one of the three and is associated with MTC, pheochromocytoma, marfanoid habitus, ganglioneuromas of the mouth and gut, and delayed puberty (Mulligan, 2014). Mutations that are associated with MEN2B are M918T and A883F, which flank the activation loop. It is hypothesized that such mutations could cause conformational changes and increase ATP binding by ten-fold and increase substrate recognition (Gujral et al., 2006; Knowles et al., 2006; Songyang et al., 1995). However, structural validation has been lacking (Plaza-Menacho et al., 2014).

CHAPTER 2

Materials and Methods

2.1 Protein Structure Datasets

2.1.1 Protein Folds

The set of protein-monomers used in this work are models targeted for refinement, referred to as TR. These targets originate from the Critical Assessment of protein Structure Prediction (CASP) experiments in their 11th and 12th round for which the reference crystal structure is available. The models are initial predictions from participating predictor groups in the template based modelling (TBM) and free modelling (FM) categories with the the aim to further improve the model quality by refinement. From all TBM or FM submissions, the model with the highest GDTTS is selected by the CASP committee for refinement. The range of starting model quality, protein length, secondary structure element composition and fold is large, which allows testing on a diverse set of proteins. In total there are 42 targets, Figure 2.1a shows three examples. The complete list of all TRs is provided in the supplemental material Table A.1 alongside PDB codes and a description. The 3D rendering of the reference crystal structure of these targets are provided in Figures A.1 and A.2 for CASP11 and CASP12 targets, respectively. These targets were used in Chapter 4 and 6.

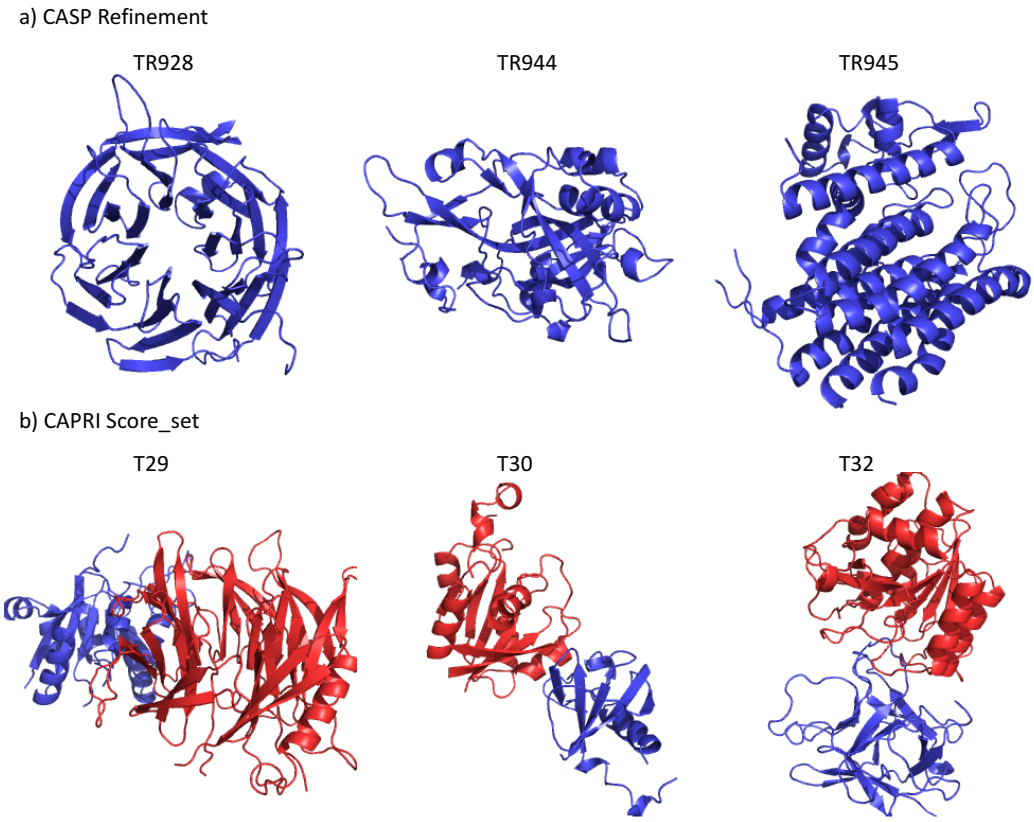

Figure 2.1: Example of CASP and CAPRI targets. The rendered 3D structures are the reference crystal structures for a) three CASP targets, b) three CAPRI targets where the colour red indicates the receptor and blue the ligand. A rendering of all used targets is provided in supplemental material Figures A.1 (CASP11), A.2 (CASP12) and A.3 (CAPRI).

2.1.2 Protein-Protein Complexes

The set of protein-protein complexes used in this work are previous targets from the Critical Assessment of Prediction of Interactions (CAPRI) experiment that were assembled in the score_set data-set (Lensink and Wodak, 2013). To be precise, this data-set contains a large variety of different protein-protein complexes where the larger and smaller entity, with respect to its number of residues, is referred to as receptor and ligand, respectively. An example of three such targets is shown in Figure 2.1b. This data-set provides a large number of solutions from a diverse set of different docking methods from participating groups. Where, depending on the target, the aim was to predict the bound complex from unbound ligand and receptor structures. In total, the data-set contains more than 19,000 docking solutions from

15 targets. However, in this work only targets which had at least one solution with acceptable or better quality were used which reduced the number to 13 (see Section 2.6.2 for a definition of quality). Table A.2 and Figure A.3 give an overview of these targets where PDB codes of their reference crystal structure and descriptions are provided.

2.2 Molecular Descriptors

2.2.1 Protein Folds

Several molecular descriptors have been used to quantify the energy or score of predicted protein monomers. All of these but one (CS_α) have been previously described in publications. The following is a short description for each descriptor:

DFIRE (Liu et al., 2004): This all-atom knowledge based statistical potential makes use of a physical state description of the ideal gas. The atom-atom potential of mean force is a function of atom type i and j and their distance r. The distance matrix was constructed for different bins ranging from 2 Å to 15 Å on a test set of 1011 non-homologous proteins (sequence identity $<$ 30%) with a resolution of $<$ 2 Å.

Reconstructed Free Energy Surface (Tribello et al., 2014): Is a reconstruction of the free energy surface (FES) from meta-dynamics. A description of meta-dynamics is provided in Section 2.4.5.

CS_α (not published): The newly introduced function CS_α combines the two energies from DFIRE and the reconstructed FES such that $CS_\alpha = (1 - \alpha)FES_N + \alpha DFIRE_N$. Where FES_N and $DFIRE_N$ are 0 to 1 normalized FES and DFIRE energies. The α parameter is set to 0.5.

DOPE (Shen and Sali, 2006): The estimated potential energy of DOPE is based on an all atom distance dependent statistical description derived from a set of 1472 crystal structures. The accuracy of the potential is improved

by explicitly modelling the reference state. This state is derived from a uniform-density sphere of finite size with a radius close to the reference native state to model non-interacting atom pairs.

DOOP (Chae et al., 2015): This energy function is based on empirical interactions of atom pair distances from a set of 954 crystal structures. Each crystal structure is decomposed into interacting fragment pairs that total to 8609 for all crystal structures. For each fragment pair, 1000 decoys are generated in order to train a neural network to learn the energy parameters of their scoring function. The authors claim that this methodology models the funnel-like energy landscape of protein folding. A drawback of this method is that it does not take into account side chain orientation dependency.

calRW (Zhang and Zhang, 2010): Is a distance dependent pair-wise statistical potential that uses an ideal random walk as the reference state. This random walk implements the freely-jointed chain (FJC) model in order to retain the chain connectivity but without modelling long range interactions between nodes. The potential was derived from 1383 non-homologous crystal structures (sequence identity $< 20\%$).

calRWplus (Zhang and Zhang, 2010): Same as calRW but integrates a side-chain orientation dependent term in the form of 20 vector pairs.

GOAP (Zhou and Skolnick, 2011): In this atomic statistical potential function the energy is calculated from the relative orientations of the planes of two interacting atoms. The reference state is described by the ideal gas state, as previously defined by DFIRE.

Molecular PDF (Eswar et al., 2008): Is the default energy function of the program Modeller (Eswar et al., 2008) and is the sum of terms describing the electrostatics, vdW force, solvation, bond, angle and dihedral energy.

2.2.2 Protein Complexes

This work used a large number of molecular descriptors to quantify the interaction, or certain aspects of it, of protein-protein complexes. An overview of the different categories is provided in Table 2.1 and the complete list with all descriptor names and their reference is shown in supplemental material Table B.1.

Table 2.1: Molecular descriptor categories. Shown are the Category name, its abbreviation, the number of molecular descriptors used in this thesis for this category and a short description. The full list of molecular descriptors is given in supplemental material Table B.1.

Category	Abbreviation	Nb. Molecular Descriptors	Description
Residue contact and distance potential	rc	34	Coarse-grained residue potentials between intermolecular residues.
Atomic contact and distance potential	ac	21	Fine grained atomic potential between intermolecular atoms.
Statistical potential constitute terms	sp	18	Knowledge based potential terms.
Composite scoring functions	cs	11	Scoring functions composed of different weighted additive terms.
Solvation energy functions	se	5	Functions describing the effect of desolvation upon protein-protein complex formation.
Hydrogen bonding	hb	3	Intermolecular hydrogen bonding.
Van der Waals and electrostatic	ve	6	Contribution of intermolecular van der Waals and electrostatics forces such as attractive and repulsive terms.
Miscellaneous	mi	11	Functions describing amino acid propensity, interface packing and change in rotational and translational entropy.

2.3 Machine Learning Algorithms

Table 2.2: Overview of machine learning algorithms. The columns show the full name of the method (Method), their abbreviation as used in this thesis (Abbreviation), the chapter these algorithms have been used in (Chapter) and the software library in which the experiments were performed in (Library).

Method	Abbreviation	Chapter	Library
Logisitic Regression	LR	6	Scikit-learn
K-Nearest Neighbours	KNN	6	Scikit-learn
Random Forest	RF	6	Scikit-learn
Extremely Randomized Trees	ERT	3	Scikit-learn
Principal Component Analysis	PCA	3	Scikit-learn
Kernal Principal Component Analysis	KPCA	3	Scikit-learn
Factor Analysis	FA	3	Scikit-learn
Recurrent Neural Networks	RNN	6	TensorFlow

In Chapters 3 and 6, several machine learning methods on classification tasks were applied. Table 2.2 provides an overview and the following subsections briefly outlines the methods. In general, the packages scikit-learn (version 0.18; (Pedregosa et al., 2011)) and TensorFlow (version 1.0; (Abadi et al., 2016)) were used to describe and train the models.

2.3.1 Logistic Regression

The logistic regression model provides a probabilistic description with respect to the desired outcome Y of a continuous input feature variable X given by:

$$p(Y = 1|X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}. \tag{2.1}$$

For the case of a binary classification problem, the output Y represents a binary variable with values 0 or 1. The class 1 is assigned if the probability of $p(Y = 1|X) >= 0.5$, otherwise 0. An example of such a probability distribution for input values X ranging from -6 to 6 is shown in Figure 2.2a. During training of the logistic regression model the two model parameter β_0 and β_1 of Equation 2.1 are optimized to yield the maximum discrimination, given the training examples with

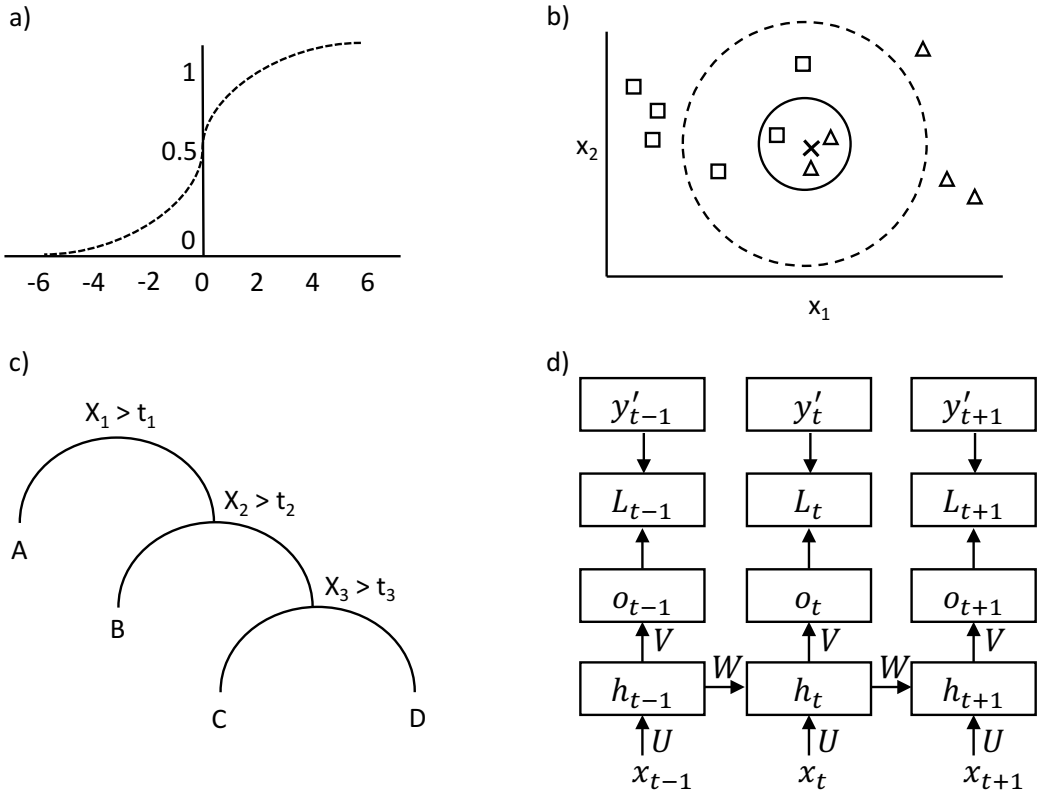

Figure 2.2: Machine learning algorithms. a) Standard logistic function $\sigma(t)$. b) K-nearest-neighbours, squares and triangles indicate data points for two classes. The cross symbol X is a new unobserved data point for which a class has to be inferred. The smaller continuous circle and the bigger dotted circle indicate the $k = 3$ and $k = 5$ neighbourhood. c) Decision tree, x and t indicate features and thresholds, respectively. The leafs A, B, C and D represent the classifications. d) Recurrent neural network, where x is the input, h the hidden state, o the output, L the loss function, y' the ground truth, V, W and U the weight matrices.

known class assignment. A method known as maximum likelihood estimation is exploited to achieve this such that

$$l(\beta_0, \beta_1) = \prod_{i:y_i=1} p(y_i|x_i) \prod_{i':y_{i'}=0} (1 - p(y_{i'}|x_{i'})) \tag{2.2}$$

is maximised. In case of multiple input features $X_1, \cdots, X_n$ the logistic regression model can be easily extended by introducing additional βs as follows:

$$p(Y = 1|X_1, \cdots, X_n) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_n X_n}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_n X_n}}. \tag{2.3}$$

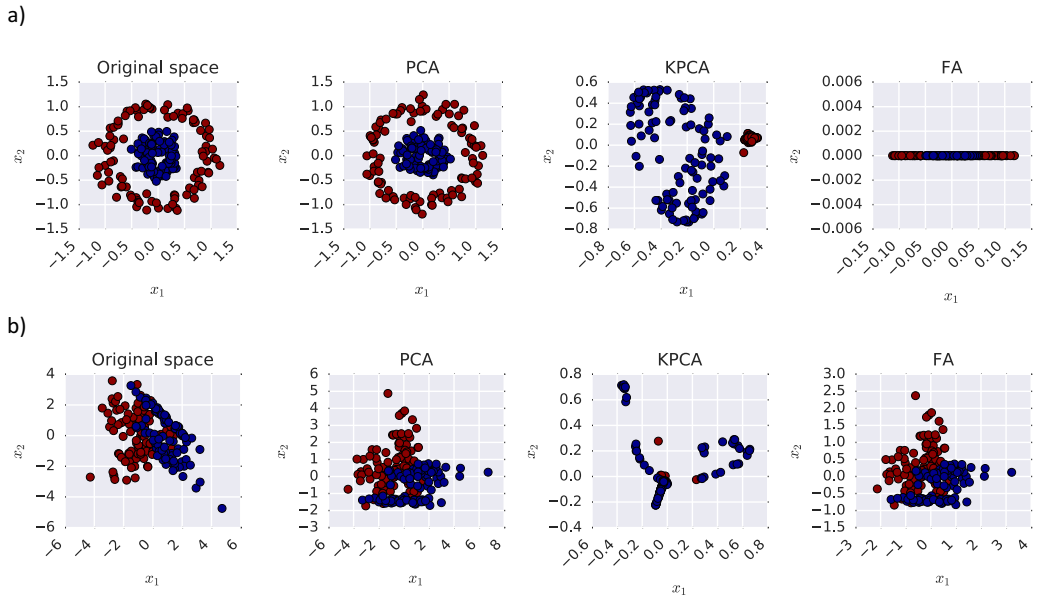

Figure 2.3: Example of PCA, KPCA and FA. Shown are the transformations for two different shapes of original feature spaces with 200 data points and two classes (red and blue) with a) circular feature space, b) random feature space.

Likewise, the training is now performed by maximizing the product of all probabilities given all βs: $l(\beta_0, \cdots, \beta_n)$. In the case of a multi-class classification problem, with K classes, where $Y = 0, 1, \cdots, K$, $K-1$ independent binary logistic regression models are trained, with one class acting as the pivot.

2.3.2 K-Nearest Neighbours

The K-nearest neighbours (KNN) classifier makes predictions for a new data-point, x_0, based on the known class assignments of the K closest data-points from the training data. A common distance metric to identify the closest training data points is the Euclidean distance given by:

$$d(x_0, q) = \sqrt{\sum_{i=0}^{n} (x_{0i} - q_i)^2}. \tag{2.4}$$

Here, the variable q is another data-point and the distance is the square-root over the sum of squared difference between x_{0i} and q_i in an n dimensional feature space.

The probability of a data-point x_0 to belong to class j can be expressed such that

$$p(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in \eta_0} I(y_i, j), \tag{2.5}$$

where η_0 is the list of the K closest neighbours and I a conditional function where

$$I(y_i, j) = \begin{cases} 1 & \text{if } y_i = j \\ 0 & \text{else.} \end{cases} \tag{2.6}$$

An example of a classification with two different K values is shown in Figure 2.2b. The figure shows a two dimensional feature space (x_1 and x_2) with 11 data-points from the training-set of which 6 have the class ”square” and 5 the class ”triangle” assigned. The new data-point, shown as the cross symbol X, is classified as a triangle with $K = 3$ and as a square with $K = 5$.

2.3.3 Random Forest

The random forest (RF) classifier is an ensemble predictor that uses a set of decision trees to assign the class k to a new data-point. A decision tree, as illustrated in Figure 2.2c, is composed of internal nodes with splitting rules of the form, for example, $X_1 > t_1$, where X_1 is a input feature variable and t_1 the so called threshold to separate examples. The terminal nodes represent the class assignments which are reached after traversing the tree starting from the root node. The construction of a decision tree, given the training data $X_1, \cdots, X_n$ with n input features and their known class assignments Y, is successively performed by finding the most discriminative splitting rules for the training data according to the Gini index G:

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}). \tag{2.7}$$

Here $\hat{p}_{mk}$ is the relative proportion of the training data with class k and the mth sub-set of the data. A small G indicates a purer split of the data, thus during training the goal is to minimize G by finding the best splitting rule.

The tree construction for a RF classifier allows for each tree to only use a random subset of the available features. For example, a frequently imposed rule is that each tree uses $\sqrt{n}$ of the features. This results in a set of trees, which are less correlated, thus reducing the variance of the model which generally leads to less over-fitting and better prediction performance on the test data (Bishop, 2006).

2.3.4 Extremely Randomized Trees

The extremely randomized tree (ERT) classifier is similar to the previously described RF (Section 2.3.3). The difference during tree construction for ERT is that the best threshold t for each splitting rule can only be chosen from a randomly generated set of threshold values. This decreases the variance of the model more but comes with an increase in bias.

2.3.5 Principal Component Analysis

Principal component analysis (PCA) is an unsupervised machine learning algorithm where only the input feature space $X_1, \cdots, X_n$ but not the class assignments Y are required. Application of principal component analysis to a n dimensional feature space allows for a projection into a lower dimensional space. The projections $Z_1, \cdots, Z_p$ represent the directions of the input feature space along which the data is most variable. For example, a feature space with $n = 10$ dimension can be projected into a principal component space with $p = 2$. Here the first component Z_1 would describe the direction with highest variability from the original input space and component Z_2 the second highest variability. The number of projections p can be $1 \geq p \leq n$. A projection into a lower dimension is especially useful when the input features are highly correlated.

A principal component, for example Z_1, is simply the linear combination of the input feature space with loading factors $\phi_{11}, \cdots, \phi_{n1}$ such that

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \cdots + \phi_{n1}X_n. \tag{2.8}$$

In order to find the first component which explains the highest variability in the data, the values for $\phi_{11}, \cdots, \phi_{n1}$ are chosen to maximize the following:

$$\underset{\phi_{11},\cdots,\phi_{n1}}{\text{maximize}} \left\{ \frac{1}{m} \sum_{i=1}^{m} \left(\sum_{j=1}^{n} \phi_{j1} x_{ij} \right)^2 \right\}, \tag{2.9}$$

with the condition that

$$\sum_{j=1}^{n} \phi_{j1}^2 = 1. \tag{2.10}$$

For the second component Z_2 the maximisation is analogous to the one described in Equation 2.9 with the additional condition that component Z_2 has to be orthogonal to component Z_1. This procedure applies analogous for all components from Z_2 to Z_p.

2.3.6 Kernel Principal Component Analysis

Kernel principal component analysis (KPCA), is by definition analogous to PCA (Section 2.3.5). The loading factors $\phi_{11}, \cdots, \phi_{n1}$ for Z_1 are chosen to maximize

$$\underset{\phi_{11},\cdots,\phi_{n1}}{\text{maximize}} \left\{ \frac{1}{m} \sum_{i=1}^{m} \left(\sum_{j=1}^{n} \kappa(x_{ij}, \phi_{j1}) \right)^2 \right\}. \tag{2.11}$$

However, instead of a linear combination, a kernel function κ is applied. In this work the radial bias function was used which is defined as

$$\kappa(x, \phi) = \exp\left(-\gamma ||x - \phi||^2\right), \tag{2.12}$$

where γ is a parameter that can be optimized. KPCA is useful for a feature space with non-linear separation between classes. Figure 2.3a shows such an example where in the original space the classes blue and red can not be separated by drawing a line. However, after transformation with KPCA (Figure 2.3a third panel from left) a separation is possible where normal PCA (Figure 2.3a second panel from

left) would produce a similar output.

2.3.7 Factor Analysis

Factor analysis (FA) is used to describe the n dimensional feature space $X_1, \cdots, X_n$ with a lower number of p unobserved factors $F_1, \cdots, F_p$ such that

$$\begin{array}{ccccccccc} X_1 & = & \phi_{1,1}F_1 & + & \cdots & + & \phi_{1,p}F_p & + & \varepsilon_1 \\ \vdots & & \vdots & & \vdots & & \vdots & & \vdots \\ X_n & = & \phi_{n,1}F_1 & + & \cdots & + & \phi_{n,p}F_p & + & \varepsilon_n. \end{array} \tag{2.13}$$

The feature vector for the first feature X_1 is described by a linear combination of the factors $F_1, \cdots, F_p$ multiplied by loading coefficients $\phi_{1,1}, \cdots, \phi_{1,p}$ and a random noise term ε_1. For example, if two factors are assumed ($p = 2$) this would equate to

$$X_1 = \phi_{1,1}F_1 + \phi_{1,2}F_2 + \varepsilon_1. \tag{2.14}$$

In the context of dimensionality reduction, the factors F_1 and F_2 would contain the data-points of the reduced feature space.

Transformations with FA can produce similar results as PCA, as shown in Figure 2.3b. However, a key difference is that factors do not have to be orthogonal to each others as compared to the principal components in PCA. This can produce, for example, a markedly different result for spherical data (Figure 2.3a).

2.3.8 Recurrent Neural Networks

In this section only a short summary of recurrent neural networks (RNN) is given. The specific model definition and training procedure as used in Chapter 6 is given in Section 6.2.1 and a broad introduction to deep learning and its applications is provided in Section 1.4.2.

A RNN consists of a directed cycle where the predictions y_t at time-point t depends on all previous and the current input $x_0, \cdots, x_t$, thus, allowing to model temporal dependencies. An illustration of such a network is shown in Figure

2.2d. Here, the input x_t to the hidden layer h_t is parametrized by weight matrix U, the input from the previous hidden state h_{t-1} to h_t by weight matrix W and the connection from h_t to output o_t by weight matrix V. For a classification task, the desired output is a probability vector y of size c, where c is the number of classes and the sum of all elements is equal to 1. However, the output o_t is not normalised. Therefore, to obtain a probability vector a softmax normalisation is applied (Goodfellow et al., 2016):

$$y_t = \text{softmax}(o_t). \tag{2.15}$$

The forward propagation equations, in order to compute the output from the first time-point $t = 0$ to the desired time-point $t = \tau$, is defined as follows:

$$a_t = b_1 + Wh_{t-1} + Ux_t, \tag{2.16}$$

$$h_t = \tanh(a_t), \tag{2.17}$$

$$o_t = b_2 + Vh_t. \tag{2.18}$$

Here, the vectors b_1 and b_2 define the bias, and at time-step $t = 0$ the hidden state h_0 is initialised with a zero matrix. The activation function for h_t is, in this case, specified as tanh; however, any other function is possible. During training the weight matrices U, W, V are learned with stochastic gradient decent (SGD), where the loss function L is defined as the cross entropy

$$L_{y'}(y) = -\sum_j (y'_j \log(y_j)), \tag{2.19}$$

with $j = 0, \cdots, c$.

2.4 Molecular Dynamics Simulations

2.4.1 The Equation of Motion

The computation of the collective motion of a system at atomic resolution can be achieved with Newton's equations of motion. The system moves under the influence of a force field that is described by a potential energy function $E(X)$, where the potential energy is dependent on the three dimensional Cartesian coordinate vector $X \in R^{3N}$ for all N atoms of the system. The motion is given by the following differential equations as described in Schlick (2010):

$$M\dot{V}(t) = F(X) = -\nabla E(X(t)) \tag{2.20}$$

$$\dot{X}(t) = V(t). \tag{2.21}$$

The velocity of the system at time t is denoted as $V(t)$, the mass M of each atom in the system is defined by a diagonal mass matrix and the superscript dot denotes the differentiation with respect to t. In these equations the force $F(X)$ can be described by the negative gradient vector of the systems potential energy at time t. The gradient is defined such that:

$$\nabla E(X)_i = \frac{\partial E(X)}{\partial \alpha_i}, \tag{2.22}$$

where $E(X)_i$ denotes the gradient for each atom in all three dimensions x, y, z written as α_i with $i = 1, \cdots, 3N$. A trajectory of length n of the systems motion is generated by numerically integrating these equations. Thus, the result is a discrete description of the system's motion given by the generated coordinate and velocity pairs $\{X^n, V^n\}$ for every time-step Δt.

2.4.2 Integration of the Equation of Motion

The integrator is an essential part of MD simulations to generate the next pair of atomic coordinates and velocities $\{X^{n+1}, V^{n+1}\}$ from the current state of the system,

$\{X^n, V^n\}$. A common scheme, and the one used in this work, is the leapfrog Verlet method (Hockney and Eastwood, 1988). In order to formulate the update scheme the acceleration of the force denoted as $\tilde{F}$ has to be defined

$$\tilde{F}(X(t)) = M^{-1}F(X(t)) = -M^{-1}\nabla E(X(t)), \tag{2.23}$$

which is the force scaled by the inverse mass matrix, M^{-1}. Given this, the leapfrog scheme can be written as:

$$V^{n+1/2} = V^{n-1/2} + \Delta t\tilde{F}^n, \tag{2.24}$$

$$X^{n+1} = X^n + \Delta t V^{n+1/2}, \tag{2.25}$$

$$V^{n-1/2} = \frac{(X^n - X^{n-1})}{\Delta t}. \tag{2.26}$$

In this set of equations the velocity is computed at half-steps, $V^{n+1/2}$, and the coordinates of the system at full-steps, X^{n+1}. From the above, the velocities at full-steps can be computed by

$$V^n = V^{n-1/2} + \frac{\Delta t}{2}\tilde{F}^n. \tag{2.27}$$

2.4.3 The Particle Mesh Ewald Method

The computation of the non-bonded interactions such as the electrostatic contribution to the energy is computational expensive. This is due to the evaluation of all atom pairs, that results in a complexity of $O\left(N^2\right)$, where N is the number of atoms. An approximation of all non-bonded interactions by the Ewald method can reduce the complexity to $O(N\log N)$ (Schlick, 2010).

For a system with N atoms, where each atom has a charge q a periodic array of replicated systems with length L is created. The long-range nature of the electrostatic interactions makes it necessary to include the interaction energy of all

replicated systems. Thus the interaction energy is (Rapaport, 2004):

$$U_{qq} = \frac{1}{2}\sum_{\boldsymbol{n}}{}' \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{q_i q_j}{|r_{ij}+L\boldsymbol{n}|} \tag{2.28}$$

where q_i and q_j indicate the charge of atom i and j respectively. This represents the sum over all vectors $\boldsymbol{n}$ where the prime indicates that self interaction with $i = j$ are omitted.

Successively , assuming charge neutrality $\sum_j q_j = 0$, the Ewald method changes this replica sum into sums over concentric spherical shells (Rapaport, 2004):

$$\begin{aligned} U_{qq} = \sum_{i<j\leq N} q_i q_j \Bigg[& \sum_{\boldsymbol{n}}{}' \frac{\boldsymbol{erfc}(\alpha|r_{ij}+L\boldsymbol{n}|)}{|r_{ij}+L\boldsymbol{n}|} \\ & + \frac{1}{\pi L}\sum_{n\neq 0}\frac{1}{n^2}\exp\left(-\frac{\pi^2 n^2}{\alpha^2 L^2}+\frac{2\pi i}{L}\boldsymbol{n}\cdot r_{ij}\right)\Bigg] \\ & + \frac{1}{2}\left[\sum_{n\neq 0}\left(\frac{\boldsymbol{erfc}(\alpha L n)}{Ln}+\frac{1}{\pi L n^2}\exp\left(-\frac{\pi^2 n^2}{\alpha^2 L^2}\right)\right)-\frac{2\alpha}{\sqrt{\pi}}\right]\sum_{j=1}^{N} q_j^2 \\ & + \frac{2\pi}{3L^3}\left|\sum_{j=1}^{N} q_j r_j\right|^2 \end{aligned} \tag{2.29}$$

where

$$\mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}}\int_x^{\infty} e^{-t^2}dt \tag{2.30}$$

is the error function and r_{ij} the distance between two atoms i and j. A common value for parameter is $\alpha = 5/L$ (Rapaport, 2004).

2.4.4 System Set-up

The system set-up for a MD simulation is composed of three parts: initialisation, equilibration and production simulation. In the following description a brief overview is given of the steps involved. The actual protocol for each experimental set-up for results described in Chapters 4, 5, 7 is given in their respective method

sections.

Initialisation The initial coordinates of a protein system are given by the experimental crystal structure. However, often these structures are not atom complete and atoms in side-chains or whole segments can be missing. In these cases molecular modelling software has to be applied to complete the structure. Another important aspect for simulations is the definition of the simulation box in which the protein system is placed together with water molecules and counter ions to neutralize the charge. The box should be big enough to avoid artifacts from periodic boundary conditions but not too big in order to reduce the computational cost. To make sure that the simulation starts from a minima in the potential energy landscape an energy minimisation has to be performed.

Equilibration The function of the equilibration is to stabilize the system before the production simulation can start. In order to start of the equilibration, initial velocities, V^0, for each of the N atoms in the systems have to be generated according to a Maxwell-Boltzmann distribution. These are $3N$ random numbers, one each for the three dimensions, from a Gaussian distribution multiplied by $\sqrt{2kT/m}$ (Fehske et al., 2007). During the equilibration, stability is reached when both the potential and kinetic energy of the system obtain convergence and fluctuation around their mean value is observed (Schlick, 2010).

Production Simulation In the production run the system is completely unrestrained and data concerning the collective motion of the atoms is collected and used for analysis.

2.4.5 Well-Tempered Metadynamics

Metadynamics is an enhanced sampling method which allows a faster and more directed exploration of different conformational states compared to classical MD simulations. Also, a reconstruction of its free energy surface (FES) as a function

of few selected degrees of freedom is possible, which are refereed to as collective variables (CVs).

In this method a bias potential B is added to the normal energy potential E of the system such that

$$B(s,t) = \Delta T \ln\left(1 + \frac{\omega N(s,t)}{\Delta T}\right), \tag{2.31}$$

where the bias is dependent on the CV definition s and its histogram, $N(s,t)$ of its variable s. The variable ω is the energy rate and ΔT is the temperature change. This discourages the sampling of frequently visited configurations of s, since B is a monotonic function (Barducci et al., 2008).

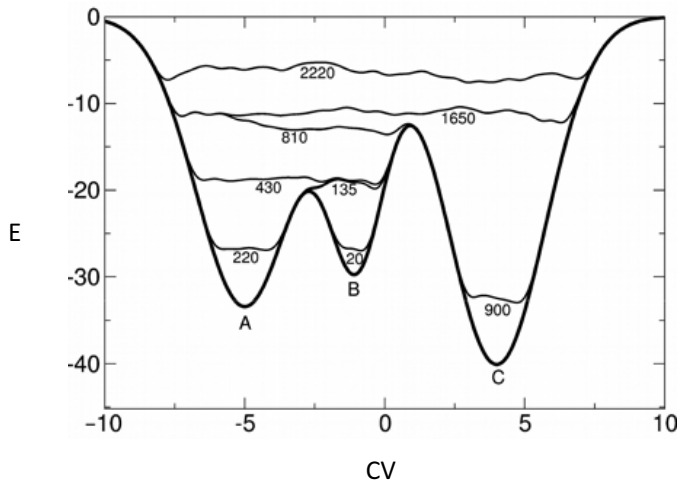

Figure 2.4: Gaussian addition in metadynamics. Example of Gaussian addition to the potential energy function E shown as a bold line with three minim A, B and C. The progressive filling is visualized by thin black lines and the numbers indicate the time of the filling in ps. Figure reproduced from Barducci et al. (2011). Permission to reproduce this Figure has been granted by John Wiley and Sons Inc.

The effect of bias-potential $B(s,t)$ on the energy landscape is shown in a schematic illustration in Figure 2.4. In this example the simulation starts in minima B of the FES of CV s. As the simulation progresses Gaussian additions of the bias potential are added to flatten the energy landscape and allowing the faster exploration of additional minima. The next minima, A, is accessible due to the

addition after 135 ps, where further additions allow the crossing after 810 ps to minima c.

2.5 Performance Measures and Significance Tests in Machine Learning

The following is a list of metrics that have been used in this work to quantify the machine learning model performance:

True positive (TP): the number of correct positive predictions by a classifier.

False positive (FP): the number of incorrect positive predictions by a classifier.

True negative (TN): the number of correct negative predictions by a classifier.

False negative (FN): the number of incorrect negative predictions by a classifier.

Accuracy: The accuracy measures the relative proportion of correctly predicted true results such that

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{2.32}$$

Recall: is a measure of the true positive rate, or hit rate. This value is maximized by reducing the FN such that

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{2.33}$$

Precision: is a measure of the ratio between correct TP predictions and incorrect FP predictions. This metric is maximized by reducing the number of FP such that:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{2.34}$$

F1-score: is a measure of the harmonic mean between recall and precision given by

$$\mathrm{F1} = 2 \times \frac{\mathrm{precision} \times \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}. \tag{2.35}$$

Pearson product momentum correlation coefficient (PPMCC): is calculated to quantify the correlation of two variables, x and y, such that

$$\mathrm{PPMC} = \frac{\sum_{i=1}^{n}(x_i - \tilde{x})(y_i - \tilde{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \tilde{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \tilde{y})^2}}, \tag{2.36}$$

where the sample mean $\tilde{x}$ (analogous for $\tilde{y}$) is defined as

$$\tilde{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{2.37}$$

2.6 Model Quality Measures for Protein Monomers and Dimers

2.6.1 Protein Monomers

The following list provides a description and definition of the different quality assessment metrics used for protein folds:

RMSD: The root mean square deviation quantifies the disagreement of the predicted model to the reference structure. A lower value means better. The definition is such that

$$\mathrm{RMSD}(v,w) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}((v_{ix} - w_{ix})^2 + (v_{iy} - w_{iy})^2) + (v_{iz} - w_{iz})^2)}, \tag{2.38}$$

where v,w are the set of atom coordinates for the model and reference structure, respectively. Where v_{ix} defines the x coordinate of the i*th* atom from set v. The convention for the 2 other dimensions and w are analogous. Usually, an optimal superimposition of v to w is performed prior to RMSD

calculation.

GDTTS: The global distance test total score is a model quality metric that evaluates quality based on the percentage of residues under different distance cutoffs given by

$$\text{GDTTS} = \frac{\text{GDT}_{P1} + \text{GDT}_{P2} + \text{GDT}_{P4} + \text{GDT}_{P8}}{4}, \tag{2.39}$$

where P_n is the percentage of residues below distance cutoff n in Å with respect to a reference structure; higher means better.

GDTHA: The global distance test high accuracy is a model quality metric similar to GDTTS but with more stringent distance cutoff values. The function is defined by

$$\text{GDTHA} = \frac{\text{GDT}_{P0.5} + \text{GDT}_{P1} + \text{GDT}_{P2} + \text{GDT}_{P4}}{4}. \tag{2.40}$$

2.6.2 Protein Dimers

The quality of predicted protein-protein complexes was assessed by a number of metrics which are defined as follows:

FNAT: The fraction of native contacts quantifies the relative number of correctly predicted residue-residue contacts between a receptor and a ligand as observed in the reference crystal structure. Where a residue-residue contact is defined as any of their atoms within a distance less than 5 Å. Values can range from 0, no correctly predicted contact, to 1, all contacts are correctly predicted.

LRMSD: The ligand root mean square deviation quantifies the translational, rotational and conformational deviation of the predicted ligand model to the reference model. The RMSD between predicted ligand position and reference ligand position is computed after optimally superimposing the receptor of the predicted complex to the reference model. The superimposition as well as

the RMSD calculation are based on Cα-atoms. This definition is slightly different from the standard CAPRI definition of LRMSD, which includes all backbone atoms (Lensink and Wodak, 2014).

IRMSD: The interface root mean square deviation describes the conformational difference at the receptor-ligand interface between the predicted model and the reference model. The set of interface atoms are given by observed residue-residue contact in the reference crystal structure. Here, a residue in the ligand is in contact with the residue in the receptor if any of their atoms has a distance $< 10\,\text{Å}$. The IRMSD calculation is based on Cα-atoms only and interface atoms of the predicted and reference model are first optimally superimposed.

CAPRI quality: The CAPRI quality is a categorical variable to specify the predicted model correctness based on an evaluation function with terms FNAT, LRMSD and IRMSD. The best possible quality assignment is ”high”, followed by ”medium”, ”acceptable” and ”incorrect”. The assignment of these quality classes for the docked solutions in the score_set data-set was directly taken from their annotation. Table 2.3 summarises the definition for each CAPRI quality class as defined by Lensink and Wodak (2014). Note, that the definition from Lensink and Wodak (2014) does not cover all edge cases. For example, a case such as FNAT=0.6, LRMSD$= 2$ and IRMSD=0.75 yields no class assignment.

Table 2.3: CAPRI quality as defined in Lensink and Wodak (2014). First column defines the name of the quality assignment where the letter in brackets indicates the abbreviation. Each row represents one rule (columns FNAT, LRMSD and IRMSD), and the quality for a predicted complex is determined by application of the rules from top to bottom till it evaluates to true.

Rank	**FNAT**		**LRMSD (Å)**		**IRMSD (Å)**
High (H)	$x \geq 0.5$	AND	$x \leq 1.0$	AND	$x \leq 1.0$
Medium (M)	$x \geq 0.5$	AND	$x > 1.0$	AND	$x > 1.0$
OR	$0.3 \leq x < 0.5$	AND	$(x \leq 5.0$	OR	$x \leq 2.0)$
Acceptable (A)	$x \geq 0.3$	AND	$x > 5.0$	AND	$x > 2.0$
OR	$0.1 \leq x < 0.3$	AND	$(x \leq 10.0$	OR	$x \leq 4.0)$
Incorrect (I)	$x < 0.1$	OR	$x > 10.0$	AND	$x > 4.0$

CHAPTER 3

A Machine Learning Approach for the Identification of Near-Native Binding Sites of Protein-Protein Complexes

3.1 Introduction

Experimental structure determination of protein-protein interactions has greatly improved our understanding of cellular processes (Jones and Thornton, 1996; Nooren and Thornton, 2003). Current experimental methods, such as X-ray crystallography and electron microscopy (EM), only provide a slow accumulation of new experimental evidence and do not allow for high throughput (Wang et al., 2015; Marsh and Teichmann, 2015). In the foreseeable future, given the large number of proteins and knowledge of the protein interaction space, a completion of this protein-protein interaction space by purely experimental methods would appear to be intractable. Instead computational protein docking is seen as the method of choice to complete missing interactions (Mosca et al., 2013). However, current state of the art methods are still not capable of routinely succeeding at this task. There are two inter-wined problems to solve, firstly, methods are

required to efficiently sample the conformational space of the interacting proteins to accommodate conformational transitions from unbound to bound states (Park et al., 2015; Vajda et al., 2013); and secondly, methods have to reliably rank the docked poses from currently thousands of generated solutions to identify ensembles, also known as clusters, or single poses that resemble native like binding. The work discussed in this chapter focuses on the correct identification of near-native clusters. This is of importance since the description of the interaction of two proteins at room temperature is physically more accurate described by an ensemble of binding modes than single crystal structure snapshots. It has been shown that there are cases in which the interaction is so diffuse that no single crystal structure snapshot gives a reliable representation of the bound state (Hamp and Rost, 2012).

There are a number of potentials that have been developed for the identification of protein-protein interactions. However, all of them suffer from a high rate of false positive predictions, whereby incorrect binding modes are highly ranked. In this chapter a novel method is presented that is based on the hypothesis of the importance of the encounter complex to identify the true-positive binding site. Essentially, the method identifies local binding site by clustering and combines it with localized SwarmDock (Moal and Bates, 2010; Torchala et al., 2013) enrichment. These enriched clusters are described by 109 molecular descriptor distributions and used as input features to train a classifier to distinguish near-native from incorrect cluster with a pair-wise learning strategy. At the time of writing, this is the first method of its kind to rank clustered poses via a machine learning approach. The training, testing and benchmarking of the method is based on the score_set dataset of docked protein-protein complex decoys (Lensink and Wodak, 2014). The models in this set originate from the Critical Assessment of PRediction of Interactions (CAPRI) experiment where protein-protein docking and ranking methods are evaluated in blind predictions (Lensink and Wodak, 2013; Janin, 2010).

In the following result sections an analysis of the 109 molecular descriptors with respect to their co-linearity and their power to discriminate near-native from incorrect clusters is presented. The results show that a reduced set of

molecular descriptors and their features is most beneficial for ranking performance. Furthermore, results are presented that show how feature space transformations based on principle component analysis (PCA) and factor analysis (FA) can help to improve the top 1 and top 5 success rate. Finally, a discussion is provided that outlines the challenges for a machine learning method based on clustered data and explains the unique properties and advantages of the solution. Additionally, the physical plausibility of the statistical model is discussed and limitations and future optimizations are presented.

The contents of this chapter are to a great part based on the work already published in Pfeiffenberger et al. (2017). The co-author Moal, I. H. performed the hierarchical-clustering of the molecular descriptors correlation matrix shown in Figure 3.3 and Bates, B. A. performed the cluster enrichment as described in Section 3.2.5.

3.2 Methods

3.2.1 Overview

The method presented here combines localized enrichment of clusters with additional solutions and training of a supervised learning algorithm to distinguish near-native from incorrect clusters. A schematic overview of this method is presented in Figure 3.1. During the training procedure of the binary classifier the predictor is optimized to predict whether $\text{LRMSD}(\text{cluster}_\text{n}) < \text{LRMSD}(\text{cluster}_\text{m})$. Here, the classifier learns from a set of pairwise cluster-comparisons where each comparison is described by 1092 features. Applying this classifier exhaustively to all possible pairwise combinations of clusters for a target produces a ranking where the best cluster has the highest number of predicted $\text{LRMSD}(\text{cluster}_\text{n}) < \text{LRMSD}(\text{cluster}_\text{m})$ and the worst cluster the least number.

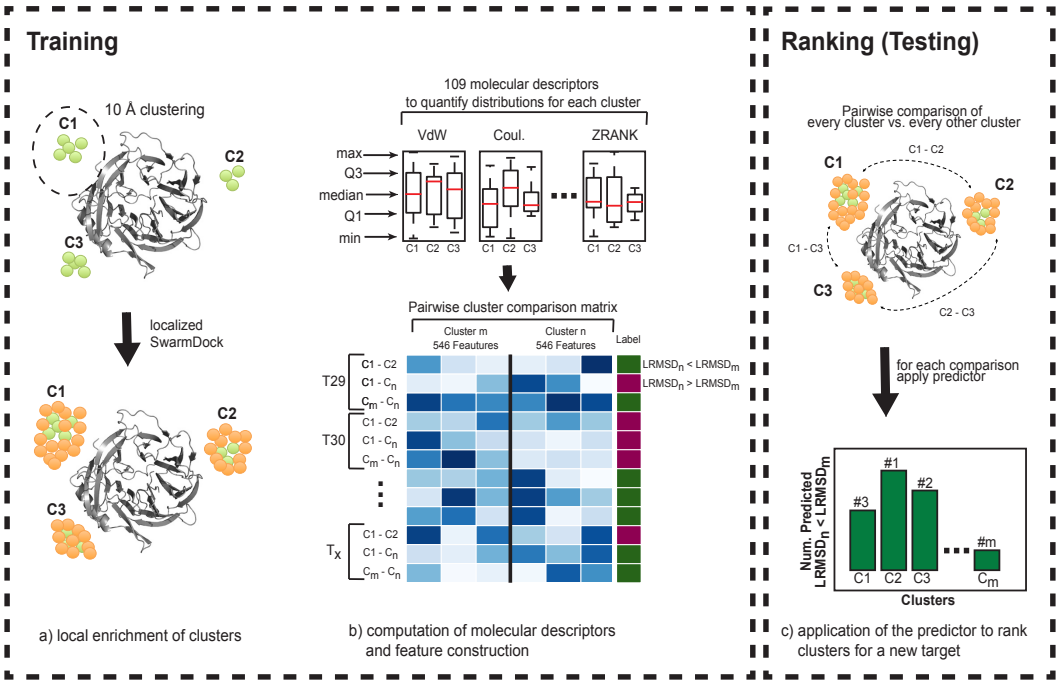

Figure 3.1: Schematic overview of the cluster ranking method. (a) decoys are clustered with a 10 Å cutoff and clusters are enriched with additional solutions with localized SwarmDock runs. Green and orange spheres around the receptor (grey) represent the centre of mass of ligand positions before and after enrichment, respectively. (b) For each model of a cluster 109 molecular descriptors are computed and grouped by cluster to quantify the protein-protein interaction. These distributions are characterized by five distribution points; minimum (min), first quantile (Q1), median, third quantile (Q3) and maximum (max) which represent the features of the supervised learning algorithm. Finally, a matrix is generated which compares all possible combinations of clusters for each target to train a binary classifier where $\text{LRMSD}_\text{n} < \text{LRMSD}_\text{m}$ produces label 1 otherwise 0. (c) To rank clusters for a new target the classifier is applied to all possible cluster comparisons. Counted is the number of times a cluster was predicted to have a lower LRMSD compared to another cluster. Ranking is based on descending order where the cluster with the highest number is ranked first and the cluster with the lowest number is ranked last. Permission to reproduce this Figure has been granted by John Wiley & Sons, Inc.

3.2.2 Dataset

The method for cluster ranking was trained and tested on protein-protein complexes from the score_set dataset which contains a large number of decoys generated from an variety of docking algorithms. In particular, the following 13 targets were used: T29, T30, T32, T35, T37, T39, T40, T41, T46, T47, T50, T53 and T54. The other two remaining targets, T36 and T38 were removed because of missing acceptable, medium or high-quality models according to the CAPRI assessment criteria. Furthermore, the crystal structure for T40 reports two possible binding sites for its ligand (see PDB 3E8L). The ligand binding position observed in chain C is denoted as T40a and for chain B as T40b. Table 3.1 gives an overview of all targets and lists the number of near native as well as incorrect solutions. Additionally, the targets T37 and T50 were randomly chosen as a hold-out set in order to test ranking performance. Hence, data from these two targets is not present in any fold of the cross-validation.

The initial dataset contained a number of models with steric-clashes (i.e. overlapping van der Waals radii between two atoms) between receptor and ligand atoms. These models were removed in order to avoid training a classifier on physically incorrect solutions. Furthermore, models for each target were stratified by modelling missing side-chains with SCWRL (Krivov et al., 2009) and by truncating receptor/ligand chains to remove residues not shared by all models.

3.2.3 Model Assessment Measures

The model quality of a receptor-ligand complex is quantified by computing the fraction of native contacts (FNAT), interface root mean square deviation (IRMSD) and the ligand root mean square deviation (LRMSD) Méndez et al. (2003, 2005).

The FNAT metric is computed on all pairs of receptor and ligand residues if any of their heavy atoms is within 5 Å. The expresses the ratio of correctly predicted residue-residue contacts in the model. This metric can range from 1.0 to 0.0 where a value of 1.0 denotes a perfect prediction where all contacts in the model are correctly predicted and a value of 0.0 denotes a model where no predicted

Table 3.1: CAPRI-Targets Overview. The total number of models with high, medium, acceptable and incorrect quality in the score_set dataset are shown. The last column indicates the number of clusters for each target. Numbers in brackets indicate the number of models or clusters after removing solutions with steric clashes. Permission to reproduce this Table has been granted by John Wiley & Sons, Inc.

Target	Total	High	Medium	Acceptable	Incorrect	Clusters
T29	2083	2	78	87	1916	925
	(1773)	(2)	(72)	(70)	(1629)	(802)
T30	1343	0	0	2	1341	741
	(1106)	(0)	(0)	(2)	(1104)	(639)
T32	599	0	3	12	584	224
	(572)	(0)	(3)	(12)	(557)	(217)
T35	499	0	0	3	496	198
	(467)	(0)	(0)	(2)	(465)	(193)
T37	1500	11	46	42	1401	629
	(1112)	(8)	(34)	(34)	(1036)	(500)
T39	1400	0	3	1	1396	465
	(1261)	(0)	(3)	(1)	(1257)	(440)
T40(a/b)	2180	193	206	189	1592	479
	(1886)	(176)	(163)	(149)	(1398)	(451)
T41	1200	2	120	249	829	141
	(1029)	(2)	(99)	(198)	(729)	(139)
T46	1699	0	0	24	1675	754
	(1321)	(0)	(0)	(24)	(1297)	(611)
T47	1051	278	307	26	440	84
	(988)	(278)	(301)	(20)	(389)	(82)
T50	1451	0	36	97	1318	306
	(1265)	(0)	(35)	(89)	(1141)	(284)
T53	1400	0	17	113	1270	277
	(1191)	(0)	(9)	(92)	(1092)	(260)
T54	1400	0	1	18	1381	301
	(1215)	(0)	(1)	(18)	(1196)	(285)

Table 3.2: Clusters with cutoff > 5 models. Shown for each target in the score_set are number of clusters (count), the number of models in the smallest cluster (Min.), the medium cluster size of a target (Med.) and the number of models for the largest cluster (Max.). For target T40 values are summerized for both interface a and b. Values in brackets indicate numbers after removing models with steric clashes. Permission to reproduce this Table has been granted by John Wiley & Sons, Inc.

		Clusters >5		
Target	Count	Min.	Med.	Max.
T29	61	6	9	147
	(49)	(6)	(9)	(136)
T30	26	6	8.5	50
	(24)	(6)	(7)	(50)
T32	12	6	9.5	168
	(12)	(6)	(9)	(166)
T35	14	4	8	131
	(12)	(3)	(7)	(128)
T37	55	6	9	35
	(41)	(6)	(8)	(27)
T39	50	6	9	94
	(44)	(5)	(8)	(94)
T40(a/b)	68	6	12	373
	(57)	(6)	(10)	(333)
T41	27	6	15	343
	(25)	(6)	(13)	(271)
T46	49	6	9	43
	(35)	(6)	(8)	(43)
T47	24	6	15	607
	(20)	(6)	(10)	(595)
T50	41	6	11	148
	(35)	(6)	(10)	(138)
T53	45	6	10.5	164
	(42)	(6)	(10)	(150)
T54	55	6	12	92
	(49)	(6)	(12)	(92)

contacts are correct with respect to the reference crystal structure.

The conformational difference between a model and the reference crystal structure at the binding interface of a protein-ligand complex is expressed by IRMSD. The interface atoms are defined as all heavy atoms within 10 Å between the receptor and ligand in the reference crystal structure. The IRMSD between a model and the crystal structure is calculated by first optimally superimposing the the defined model interface atoms to the crystal structure followed by calculating the root mean square deviation (RMSD).

The LRMSD quantifies the overall geometrical and conformational difference of the ligand of a docked model with respect to the reference crystal structure of the complex. The LRMSD between a model and the reference crystal structure is computed by optimally superimposing the equivalent receptor Cα atoms, followed by an RMSD calculation based on the ligand Cα atoms. This LRMSD calculation is different from the CAPRI standard where all backbone atoms (i.e. N, Cα and O) are used for both, superimposition and the calculation of the backbone deviation.

The model quality annotations used in this work are taken from the score_set dataset and are a function of backbone LRMSD, IRMSD and FNAT. A detailed description is provided in Lensink and Wodak (2014) and Section 2.6.2.

3.2.4 Clustering

The clustering for each target was performed with the GROMACS software package (Pronk et al., 2013) where the GROMOS clustering algorithm (Daura et al., 1999) with a 10 Å LRMSD cutoff was used. The GROMOS clustering algorithm is a greedy clustering technique that in each iteration tries to find the largest cluster given the LRMSD cutoff size until all solutions belongs to a cluster. This produces clusters where all members of a cluster are within 10 Å LRMSD to the centroid model. Table 3.1 gives an overview of the total number of clusters for each target.

In this work, a cluster-size cutoff of > 5 solutions was imposed. This constraint was applied for two reasons: i) to make the cluster enrichment computationally feasible with the resources available since the number of clusters can be as large

as 925 (see target T29); ii) if the docking community is able to find near-native solutions to a target it is populated by a number of solutions. However, for the two targets T35 and T39 an exception to this assumption was observed, where the near native cluster size was 3 and 5, respectively. In order to have the largest possible data-set available for testing and ranking these targets were included.

3.2.5 Cluster Enrichment

In order to gain a better understanding of the local binding energy distribution, additional solutions for each cluster were generated with SwarmDock runs. This cluster enrichment used the unbound ligand-receptor starting conformation and was performed with 250 particles, where the starting positions of each particle was limited to the 10 Å LRMSD of each cluster. To promote a more diverse set of binding modes for each cluster SwarmdDock runs were not allowed to fully converge. Therefore, the final ensemble of SwarmDock generated poses was typically less than 10 Å but not under 3 Å. Energy minimizations with CHARMM (Brooks et al., 2009) for all SwarmDock generated poses were performed in order to minimize the occurrence of clashes. If clashes were still present after this step, they were removed from the set. Here, a clash is defined as two atoms overlapping by their van der Waals radii.

The enrichment data was sub-clustered again in order to identify distinct docking poses and to make the computation of the molecular descriptors computationally tractable. The GROMOS algorithm with a cutoff of 3 Å was applied to each swarmed cluster. A ranking according to cluster-size was performed and the centroid structure of each of the top 10 most populated sub-clusters were retained. This resulted in 10 additional models, if 10 or more clusters were present.

3.2.6 Computation of Molecular Descriptors and Feature Construction

For each model within the enriched clusters a total of 109 molecular descriptors were computed. These contain descriptors from the CCharPPI server (Moal et al., 2015b) in addition to two manually computed descriptors, namely DCOMPLEX (Liu et al., 2004) and ZRANK (Pierce and Weng, 2007). DCOMPLEX is an atomic contact potential (see Section 2.2.2 for a description) and ZRANK is a composite scoring function with weighted terms (see Section 2.2.2 for a description). Essentially, all these can be categorized into residue contact and distance dependent potentials (rc), atomic contact and distance dependent potentials (ac), constituent terms of statistical potentials (sp), composite scoring functions (cs), solvation energy functions (se), and van der Waals and electrostatic potentials (ve). Table 2.1 provides an overview of the number of descriptors for each category. A detailed list of each descriptor together with its abbreviation, associated features, category and reference can be found in supplemental material Table B.1.

The standardisation of the data is performed by scaling the data points to zero mean and unit variance for each complex. The values are aggregated by cluster, providing 109 distributions (see Figure 3.1b). These cluster distributions are characterized by five distribution points: minimum (MIN), 1st quartile (Q1), median (AVG), 3rd quartile (Q3) and maximum (MAX). Additionally, the cluster size was added as a feature. Overall, this results in 546 features describing a cluster. As an example, the feature with the label C2_Q3_N_CP_D1 denotes a feature calculated for the second cluster (C2) in a cluster comparison and represents the 3rd quartile (Q3) of a normalized distribution (N) for the DECK potential (CP_D1).

3.2.7 Training, Testing and Ranking

The training of the classifier is performed by learning from an exhaustive set of pairwise cluster comparisons, where every comparison appears only once in the matrix. The comparison of cluster m and n in each training example contains 1092

features (i.e. 546 features describing one cluster, see Figure 3.1b). The resulting matrix is used to train an extremely randomized tree classifier (ERT) from the scikit-learn machine learning library (Pedregosa et al., 2011) where the training task is to correctly assign the label 1 if $\min(\text{LRMSD}_n) < \min(\text{LRMSD}_m)$ otherwise 0. Training of the ERT classifier was performed using 3000 trees where samples are bootstrapped and the gini impurity criterion is used to decide on the quality of the splits when building the trees. Out-of-bag samples were used to estimate the generalisation error. Each individual tree uses 33 features ($= \sqrt{1092}$) and has a maximum depth of 100, where the minimum sample size per leave is 1. These parameters represent empirical good values for classification tasks of tree-based classifiers (Breiman, 2001; Geurts et al., 2006). The performance of the classifier was tested with Leave-Complex-Out Cross-Validation (LCO-CV). Every fold uses data from n-1 targets for training and leaves out all training examples for the target it has been tested on, hence yielding 11 cross-validation folds.

The cluster ranking, by employing the above classifier, is based on the number of times a cluster was predicted to have a lower LRMSD than every other cluster for a target. The cluster with the highest number of assignments is ranked first and the cluster with the least assignments last (see Figure 3.1c). The cluster with the lowest LRMSD is assigned as being the correct true positive solution and is refereed to as the best LRMSD cluster or best near native cluster. These ranking results are compared to a base-line ranking protocol where the same clusters, containing the score_set and enrichment models, are compared to DCOMPLEX (Liu et al., 2004). This energy function is based on DFIRE and derives a potential energy for protein-protein interaction from inter-atomic distances at the interface. Here, clusters are ranked according to the cluster model with the lowest energy.

3.2.8 Molecular Descriptor, Feature and Classifier Performance Measures

The Mann-Whitney U-test was used to assess a molecular descriptor's power to discriminate between near-native and incorrect solutions. Additionally, the

Pearson's Product Momentum Correlation Coefficient (PPMCC) was calculated between all possible pairs of molecular descriptors to quantify their correlation. The classification performance of the different ERT classifiers is measured by recall, precision, F1-score and accuracy. The definition of these metrics is given in Section 2.5. Additionally, the relative feature importance identified by the internal ERT function is computed. Where the feature importance is based on the fraction of samples upon which a feature will have bearing.

3.2.9 Feature Space Reduction and Transformation

Three different methods were used to reduce the dimensionality of the feature space, namely: factor analysis (FA), principal component analysis (PCA) and kernel PCA with a radial basis function (KPCA). This reduction was performed incrementally where the initial 1092 dimensions were reduced to 2 dimensions in steps of 10 for FA and KPCA and in steps of 1 for PCA. For each step, the transformer of each respective method was fitted using the training data which was then applied to the test data. This was performed for every fold of the LCO-CV.

Following the transformation of the data, an ERT classifier was trained for each CV-fold and incremental step of the dimensionality reduction. Their performance on the test data was assessed with the metrics recall, precision, F1-score and accuracy. Finally, the ERT model with the best average F1-score from the LCO-CV was selected for assessing their ranking performance. Thus, three models are generated: ERT+PCA (i.e. ERT classifier trained and tested on transformed data with PCA), ERT+FA (i.e. ERT classifier trained and tested on transformed data with FA) and ERT+KPCA (i.e. ERT classifier trained and tested on transformed data with KPCA).

3.2.10 Recursive Feature Elimination

The initial feature space of 1092 is recursively reduced by 10 features in each round. In each round, the 10 features removed have the least average relative feature importance as identified by the LCO-CV. The ERT performance in each round is

assessed by the metrics recall, precision, F1-score and accuracy. The Ranking performance is assessed based on the model with the best average F1-score and this model is denoted as ERT+RFE in the following sections. Additionally, feature space transformation with FA was tested again on the reduced feature set and is named ERT+RFE+FA.

3.3 Results

The reported results in the following sections are based on all targets and T40a for metrics top 1, top 5, top 10, average rank and relative ranking improvement with respect to DCOMPLEX. Results for all targets and T40b are not reported in the text for clarity reasons, but can be seen in Table 3.3.

3.3.1 The Effect of Localized Enrichment on Near Native Clusters

The SwarmDock enrichment was analysed with respect to whether improved binding modes can be generated in the best near-native cluster. Therefore, LRMSD, IRMSD and FNAT were also computed for the enrichment models. In general, the enrichment produces solutions with a limited range in order to stay within the 10 Å boundaries of the clusters as seen in Figure 3.2.

The method has limited success in improving the quality of the models. The LRMSD improved for the targets T30, T32, T39 and T53 (4 out of 11). The best LRMSD improvement was observed for T30, with a decrease by 2.95 Å. Similarly, for IRMSD, improvements for T30, T39, T46 and T53 where seen. Here the largest improvement was observed for T30 which showed a decrease of 1.94 Å in IRMSD. For FNAT, only T30 could be improved by the moderate value of 0.045.

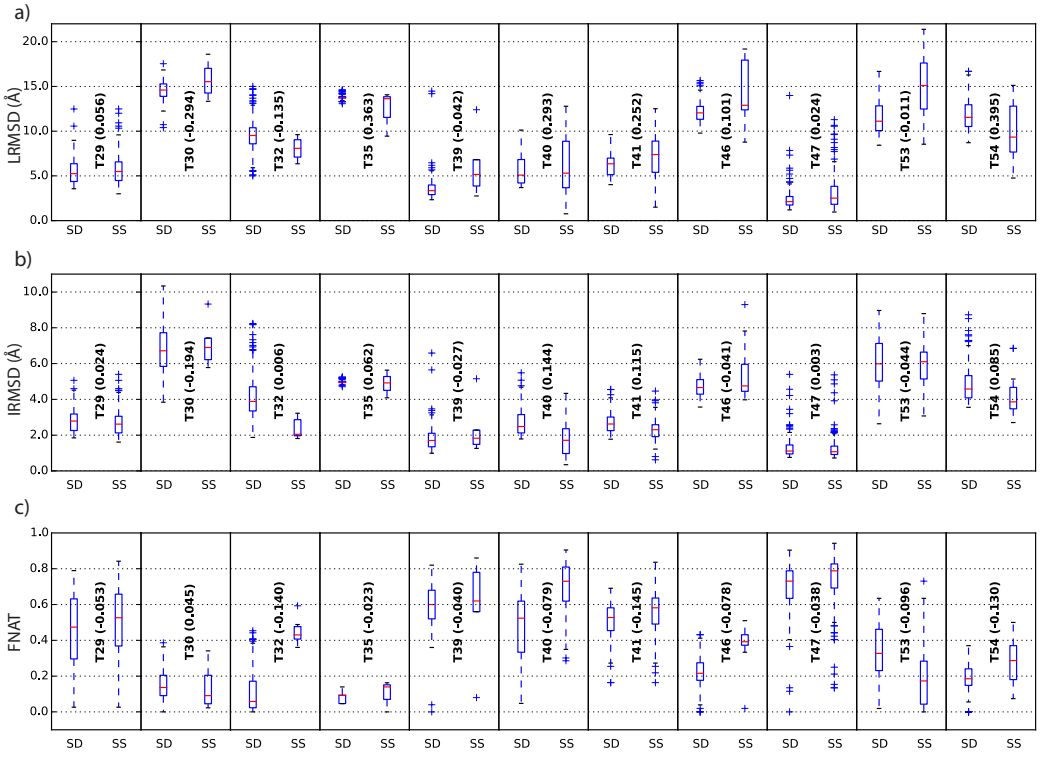

Figure 3.2: Comparison of score_set (SS) models vs. SwarmDock (SD) local enrichment models. Shown are comparisons for the cluster closest to the native binding site for each target, except for hold-out targets. Metrics shown are (a) LRMSD, (b) IRMSD and (c) FNAT. Permission to reproduce this Figure has been granted by John Wiley & Sons, Inc.

3.3.2 Molecular Descriptors, Discriminative Power and Cross-Correlation

In order to discriminate between near native and incorrect cluster, the computation of a U-test for all molecular descriptors was performed. The analysis showed that 99 out of 109 descriptors are able to produce a significant difference (p-value < 0.01) between these two groups. The top 10 descriptors are shown in Figure 3.4a. The best descriptor N_CP_TB, known as the TOBI potential (Tobi and Bahar, 2006), has a good discrimination at the 1st quartile to 3rd quartile level. However, even for this descriptor many low energy outliers within the incorrect clusters make a clear separation of the two groups hard. This is a common theme for all studied descriptors where a high number of low energy outliers makes it impossible to

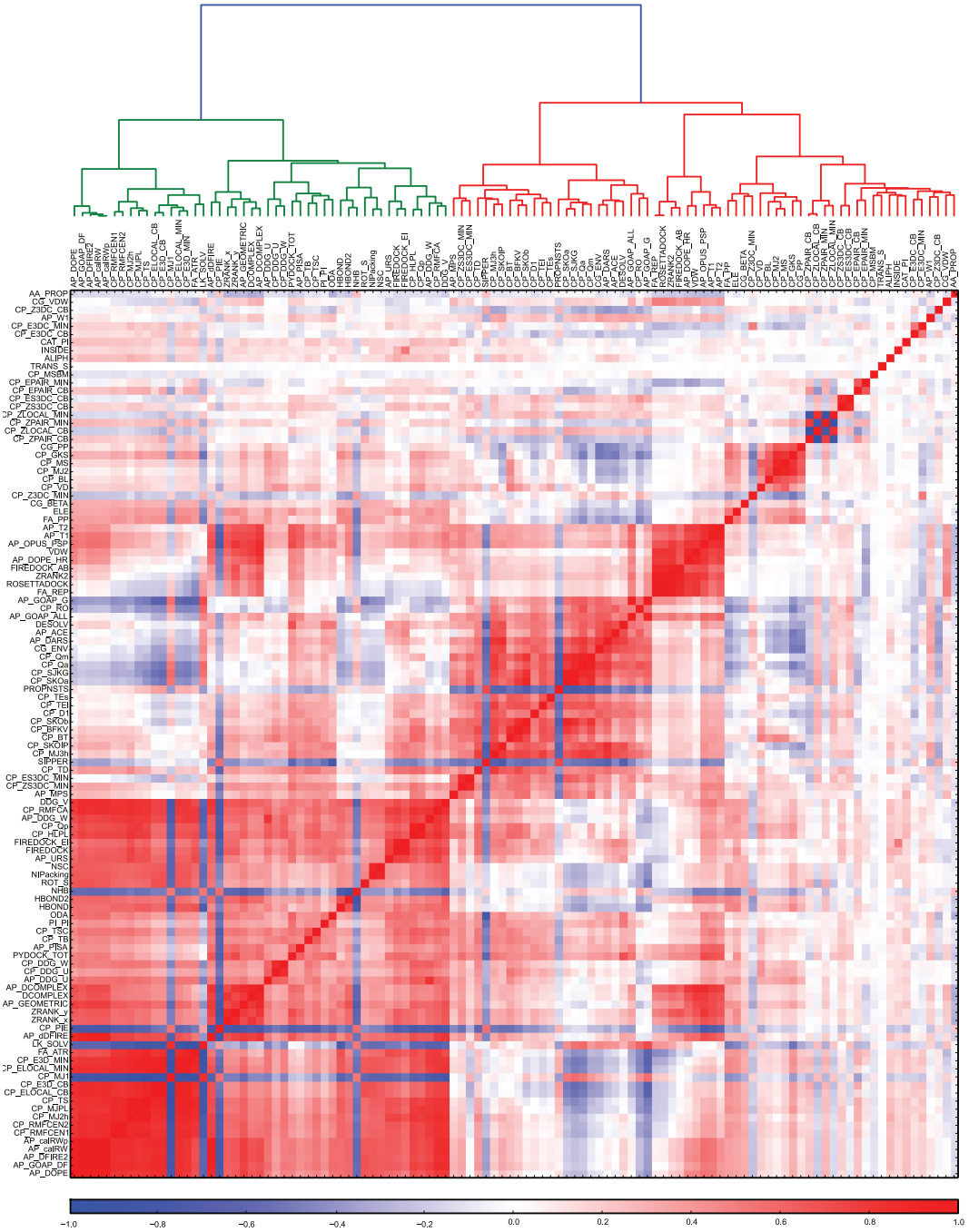

Figure 3.3: Co-linearity of all molecular descriptors. The heat-map shows the PPMCC of all pairs of molecular descriptors. Red and blue indicate high positive and negative correlation respectively. White indicates no correlation. The grouping of the molecular descriptors is based on hierarchical clustering where the distance $d = \sqrt{2(1-|PPMCC|)}$. Permission to reproduce this Figure has been granted by John Wiley & Sons, Inc.

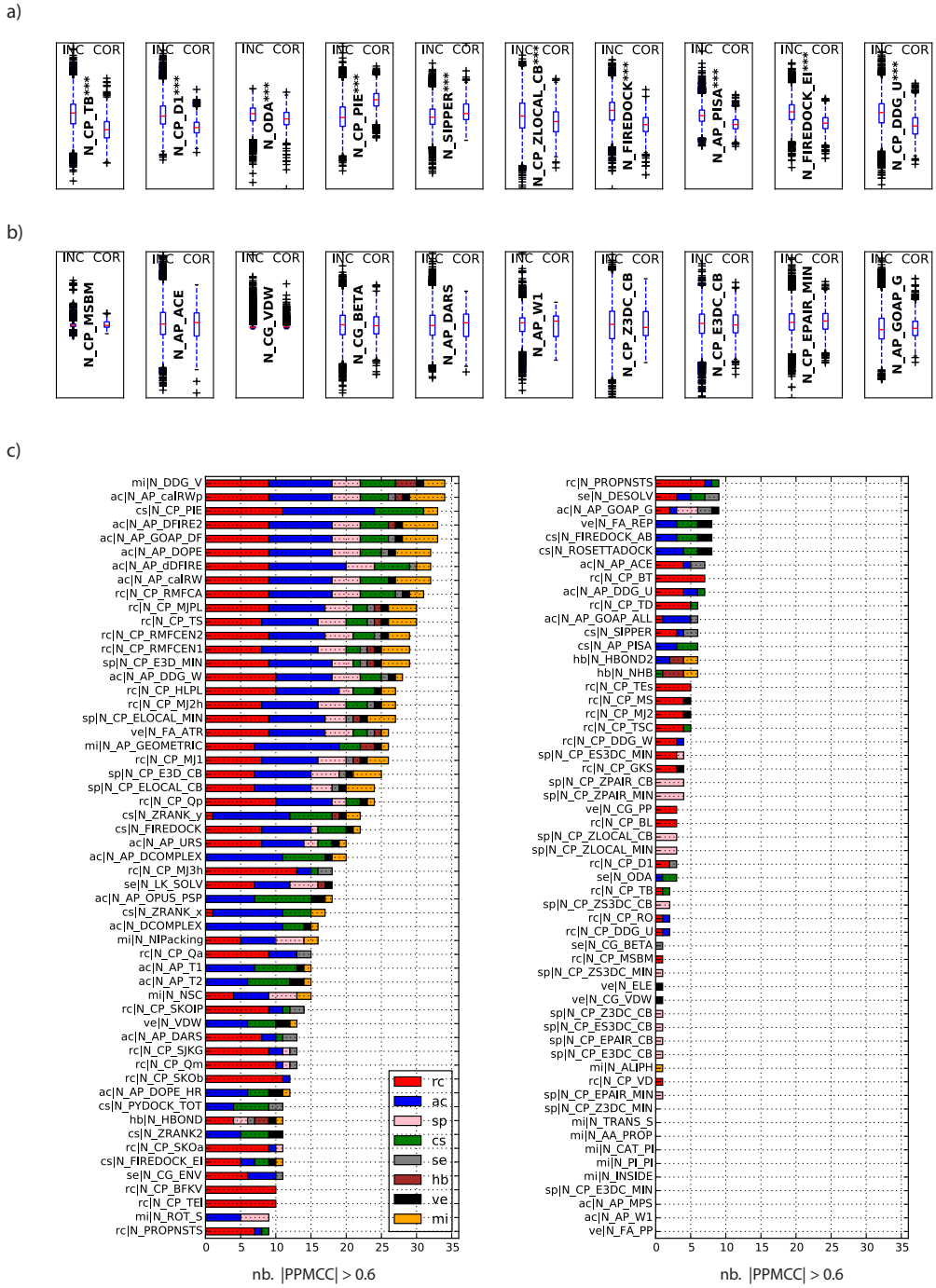

Figure 3.4: Distribution and correlation of molecular descriptors. a) and b) shows the distributions of the top 10 and bottom 10 molecular descriptors for near native/correct clusters (COR), versus clusters that contain only incorrect solutions (INC). Stars indicate p-value for U-test between groups COR and INC (***: p-value < 0.0001, **: p-value < 0.001 and *: p-value < 0.01). It can be seen that the value ranges between the groups INC and COR heavily overlap, thus, explaining to the difficulty of correctly identifying the correct binding site. c) Number of highly correlated molecular descriptors with a $|\text{PPMCC}| > 0.6$, coloured by category. Permission to reproduce this Figure has been granted by John Wiley & Sons, Inc.

clearly separate near native from incorrect clusters.

In addition to their discriminative power, descriptors were also tested for their unique information value by computing the PPMCC for all possible molecular descriptor pairs. The resulting heatmap, shown in Figure 3.3, of this computation shows a large number of positively and negatively correlated descriptors. In order to quantify the extend of correlation the number of highly correlated descriptor pairs (i.e. $|\text{PPMCC}| > 0.6$) was counted. This analysis shows that 11 descriptors have a high correlation with 30 or more other descriptors (see Figure 3.4c). The two highest correlated descriptors, N_DDG_V (Moal et al., 2015a) and N_AP_calRWp (Zhang and Zhang, 2010), are a microscopic surface energy model and an orientation dependent potential, respectively. The other descriptors in this category are either residue-contact/distance potentials or atomic-contact/distance potentials and have mostly high correlations to descriptors within their own category. Interestingly, the TOBI potential (N_CP_TB), one of the highest discriminative descriptors, has only high correlations with two other descriptors (N_CP_PIE (Ravikant and Elber, 2010) and N_CP_TSC (Tobi, 2010)). Therefore, provides highly non-redundant information, see Figure 3.4c.

Furthermore, the 10 descriptors with no statistical difference between correct in incorrect clusters (p-value > 0.01, see Figure 3.4)b, also have low numbers of high correlations to other descriptors, as seen in Figure 3.4c.

3.3.3 Ranking and Feature Performance of the Standard ERT Classifier

The results for the standard ERT classifier are presented where all 1092 features are used to train the model. Overall, this method is able to rank the best LRMSD cluster in the top 1 for 4 targets, in the top 5 for 9 target and in the top 10 for 12 targets. An average rank of 4.6 (35% percent improvement to DCOMPLEX) is achieved. For target T29 the method ranked the cluster closest to the near-native binding site first. Furthermore, a good correlation coefficient of 0.663 between predicted and actual was achieved (see Figure 3.5a). That has the effect that 9 out of the top 10 ranked

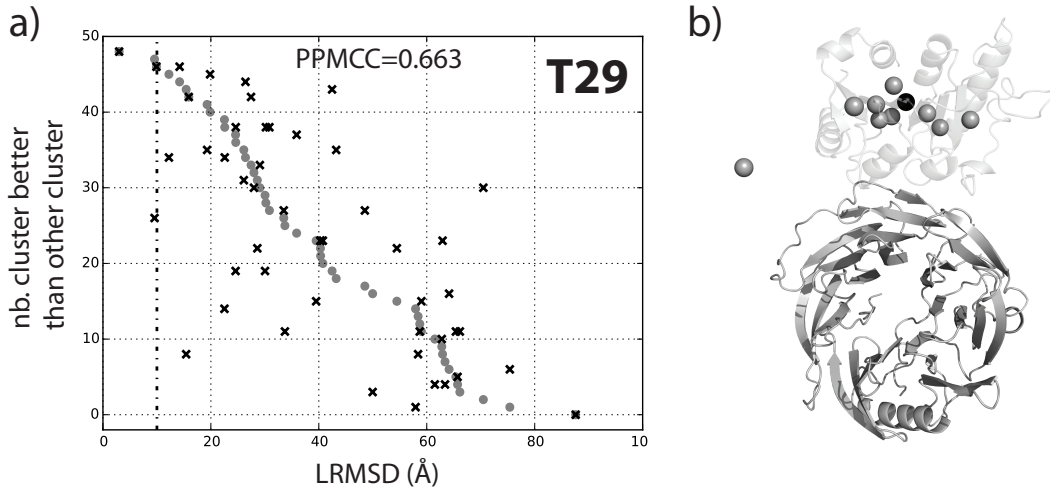

Figure 3.5: Predictions for T29 based on the standard ERT classifier. (a) the predicted number of times a cluster is better vs. all other clusters (black cross) compared to the actual values (gray dots). The LRMSD values on the x-axis are based on the cluster member with the lowest LRMSD. The bottom panel (b) shows the receptor (dark gray cartoon representation) and a sphere indicating the center of mass of the centroid model for each cluster. The top 10 ranked clusters (black: rank 1, gray: rank 2-10) are shown. The transparent cartoon indicates the observed position of the ligand from the crystal structure (PDB: 2VDU). Permission to reproduce this Figure has been granted by John Wiley & Sons, Inc.

clusters are in close proximity to the true positive binding site as shown in Figure 3.5b.

Feature importance ranges from 0.003 to 0.0007 for all 1092 features, thus no feature is dominating. Sorting the features in descending order of feature importance, as shown in Figure 3.7a, revealed a drop of the relative importance after the first 20 features. The cumulative fraction for the 200 best ranked features show that features from the categories cs (42%), se (30%), rc (21%) and sp (20%) have a high dominance whereas features from the categories hb and ve are under-employed and were first observed at rank 420 and 150, respectively (see Figures 3.7b and c).

An analysis of the top 20 features as seen in Figure 3.7d shows that these are dominated by the descriptors from the TOBI potential (N_CP_TOBI, rc), the DECK residue level distance-dependent potential (N_CP_D1, rc) (Liu and Vakser, 2011), the optimal docking area (N_ODA, se) (FernandezRecio et al., 2005), PIE score (N_CP_PIE, cs), SIPPER (N_SIPPER, cs) (Pons et al., 2011),

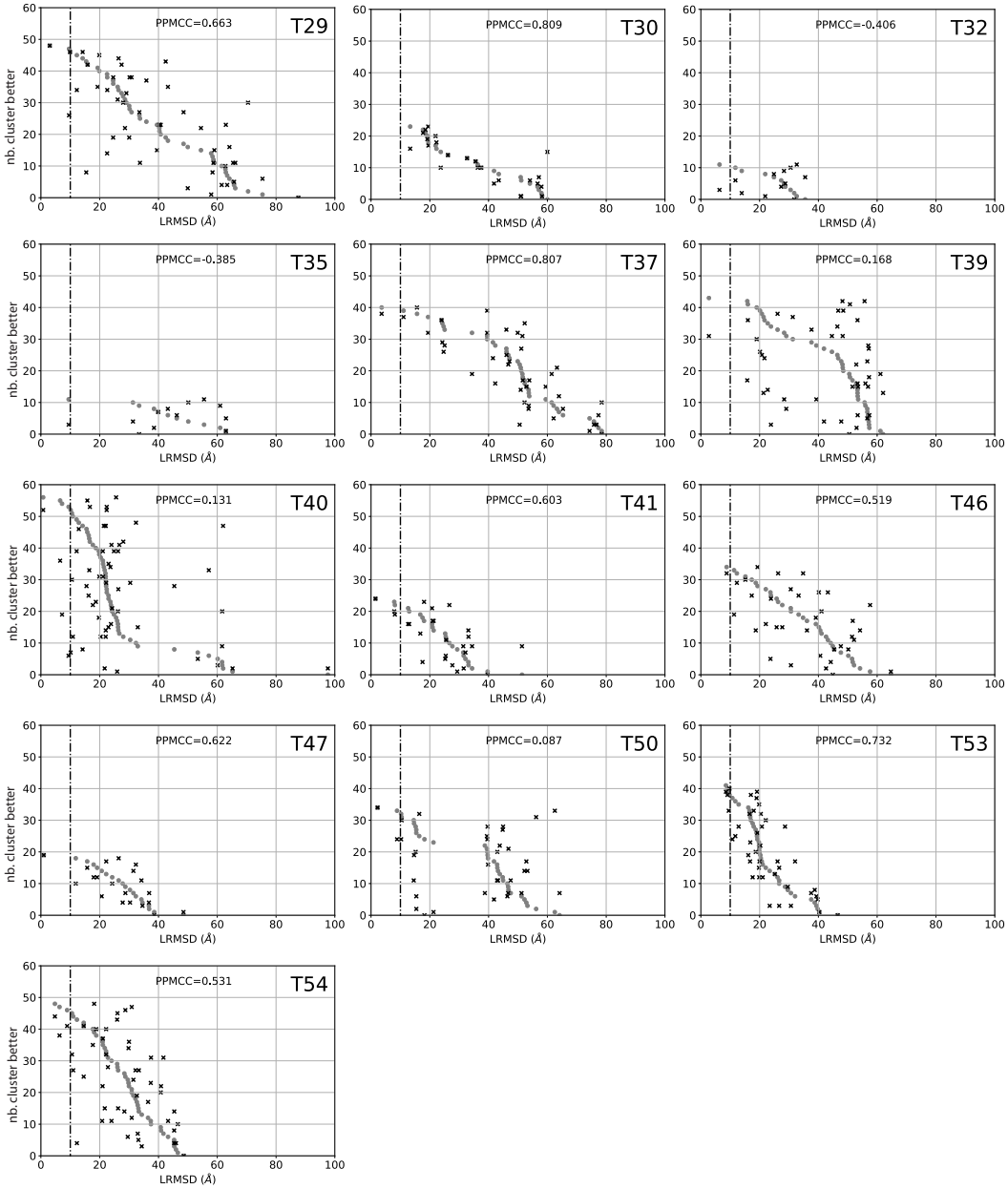

Figure 3.6: The predicted number of times a cluster is better vs. all other clusters (black cross) compared to the actual values (gray dots). The Cα-LRMSD values on the x-axis are based on the cluster member with the lowest Cα-LRMSD. Permission to reproduce this Figure has been granted by John Wiley & Sons, Inc.

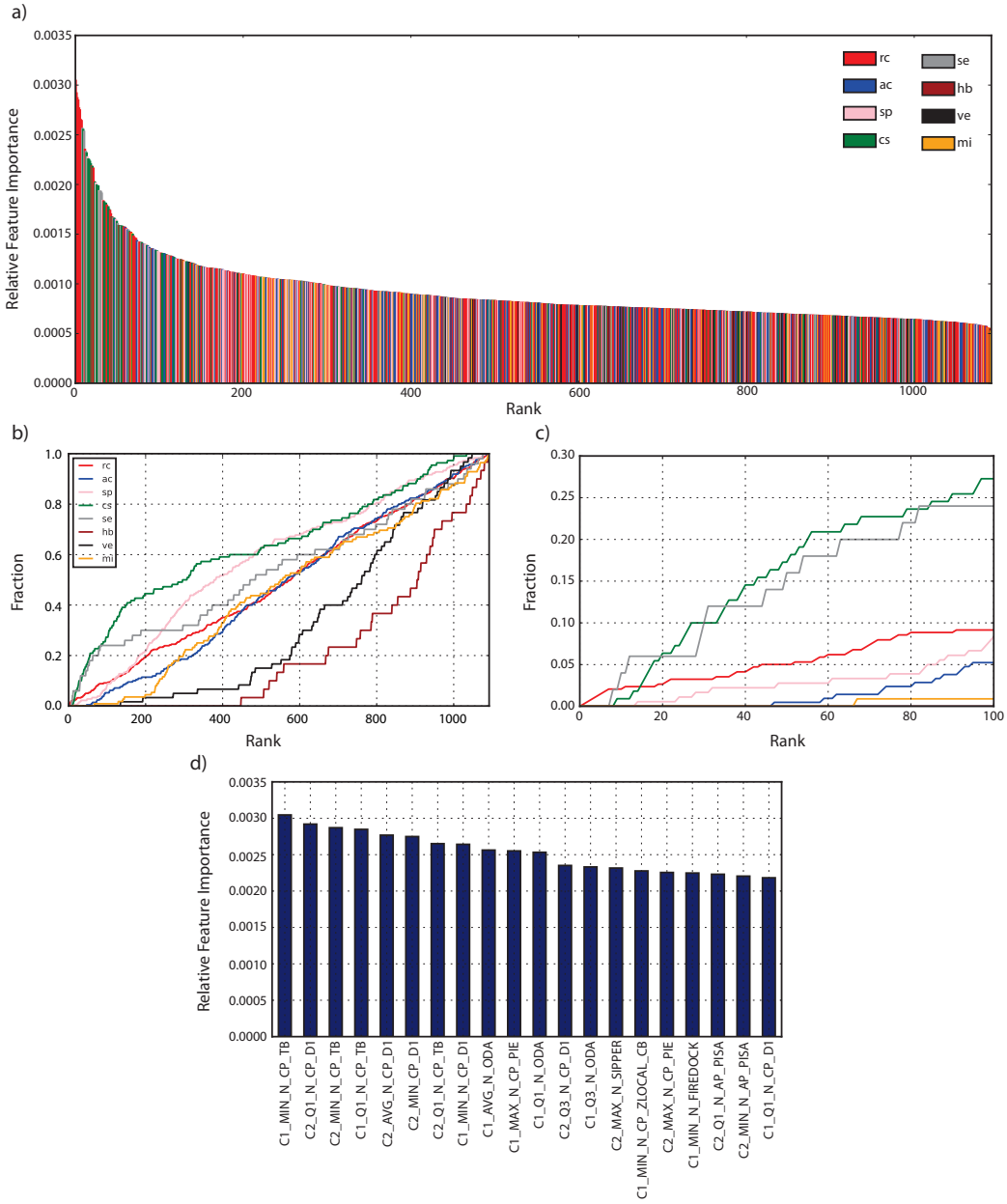

Figure 3.7: Feature importance. The top panel (a) shows the relative importance of all 1092 features colored by category. The bottom panels show the cumulative fraction of features for different categories for ranks 1 to 1092 (b) and 1 to 100 (c). Panel (d) shows the relative feature importance of the top 20 features. Permission to reproduce this Figure has been granted by John Wiley & Sons, Inc.

the w_local Z-score C_beta potential (N_CP_ZLOCAL_CB, sp) (Feliu et al., 2011), the FireDock energy function (N_FIREDOCK, cs) (Andrusier et al., 2007) and the PISA score (N_AP_PISA, cs) (Viswanath et al., 2013). The best descriptor TOBI has features at ranks 1 (C1_MIN_N_CP_TB), 3 (C2_MIN_N_CP_TB), 5 (C1_Q1_N_CP_TB) and 7 (C2_Q1_N_CP_TB). Also, the descriptor of the DECK residue level distance-dependent potential has 6 occurrences in the top 20 with ranks 2 (C2_Q1_N_CP_D1), 5 (C2_AVG_N_CP_D1), 6 (C2_MIN_N_CP_D1), 8 (C1_MIN_N_CP_D1), 12 (C2_Q3_N_CP_D1) and 20 (C1_Q1_N_CP_D1).

Table 3.3: Model performance for each target. The rank of the cluster that contained the model with the lowest LRMSD to the crystal structure are shown, referred to as best near native (NN) cluster. Number of models for this cluster with high (H), medium (M), acceptable (A) and incorrect (I) are shown along with the LRMSD of the best and centroid model in the cluster. The summary Top 1, Top 5, Top 10 shows the number of times a cluster was ranked in the respective top n category out of all 13 targets. The rows Avg. Rank and Rel. Imp. report the average rank and its relative improvement to DCOMPLEX (DC) respectively. The summary rows Top 1, Top 5, Top 10, Avg. Rank, Rel. Imp. first report values considering interface T40a and in brackets for interface T40b. The * indicates targets used in the hold-out set. Permission to reproduce this Table has been granted by John Wiley & Sons, Inc.

Target	Best NN Cluster, Nb. of Models (H/M/A/I)	Best NN Cluster, Best Model (Å)	Best NN Cluster, Centroid Model (Å)	DC	ERT	ERT+ FA	ERT+ PCA	ERT+ KPCA	ERT+ RFE	ERT+ RFE+ FA
T29	2/72/60/2	3.01	3.8	7	1	1	1	1	1	1
T30	0/0/2/7	13.32	13.32	12	8	7	7	8	8	6
T32	0/3/6/0	6.35	6.56	10	9	8	5	6	9	5
T35	0/0/2/1	9.44	9.44	12	9	12	12	12	10	12
*T37	1/14/8/0	3.61	8.3	1	3	15	10	4	4	2
T39	0/3/1/1	2.75	2.75	38	12	13	35	39	4	33
T40a	90/139/104/0	0.76	4.21	1	5	1	1	1	6	3
T40b	86/20/11/15	0.57	1.22	2	2	25	3	2	1	2
T41	2/99/170/0	1.5	5.41	3	1	1	1	1	1	1
T46	0/0/12/25	8.77	12.83	1	3	2	3	3	2	1
T47	278/301/14/2	0.96	1.38	4	1	1	1	1	1	1
*T50	0/35/85/18	2.22	6.74	1	1	1	1	1	1	1
T53	0/9/66/75	9.44	15.11	2	2	1	1	1	1	1
T54	0/1/7/8	4.76	8.72	1	5	5	5	6	3	3
Top 1				5	4	6	6	6	5	6

Table 3.3: Model performance for each target. The rank of the cluster that contained the model with the lowest LRMSD to the crystal structure are shown, referred to as best near native (NN) cluster. Number of models for this cluster with high (H), medium (M), acceptable (A) and incorrect (I) are shown along with the LRMSD of the best and centroid model in the cluster. The summary Top 1, Top 5, Top 10 shows the number of times a cluster was ranked in the respective top n category out of all 13 targets. The rows Avg. Rank and Rel. Imp. report the average rank and its relative improvement to DCOMPLEX (DC) respectively. The summary rows Top 1, Top 5, Top 10, Avg. Rank, Rel. Imp. first report values considering interface T40a and in brackets for interface T40b. The * indicates targets used in the hold-out set.

Target	Best NN Cluster, Nb. of Models (H/M/A/I)	Best NN Cluster, Best Model (Å)	Best NN Cluster, Centroid Model (Å)	DC	ERT	ERT+ FA	ERT+ PCA	ERT+ KPCA	ERT+ RFE	ERT+ RFE+ FA
				(4)	(4)	(5)	(5)	(5)	(6)	(6)
Top 5				8	9	8	9	8	9	10
				(8)	(9)	(8)	(9)	(8)	(10)	(10)
Top 10				10	12	10	10	11	13	11
				(10)	(12)	(9)	(10)	(11)	(13)	(11)
Avg.				7.2	4.6	5.2	6.4	6.5	3.9	5.4
Rank				(7.2)	(4.4)	(7.1)	(6.5)	(6.5)	(3.5)	(5.3)
Rel.					35%	27%	11%	10%	45%	25%
Imp.					(39%)	(2%)	(10%)	(10%)	(51%)	(27%)

3.3.4 The Effect of Feature Space Transformation on Prediction Accuracy

The results in Figure 3.3 show that a large number of descriptors are strongly correlated. Hence, dimensionally reduction with PCA, FA and KPCA was applied to test whether improved prediction performance can be achieved. Figure 3.8 shows that recall can be significantly improved from 0.72 (standard ERT) to 0.89 (382 dimensions), 0.92 (224 dimensions) and 0.98 (1082 dimensions) for FA, PCA and KPCA, respectively.

The accuracy of 0.62 yield by the standard ERT classifier stays largely unchanged after dimensionality reduction. However, the dimensionality could be greatly decreased to 92 (FA), 112 (KPCA) and 130 (PCA), respectively, without performance loss. To be precise, an accuracy of 0.62 was achieved for FA and KPCA and 0.63 for PCA. Similar small changes were observed for precision after feature space transfromation. Here, the precision of the standard ERT classifier with 0.62 changes to 0.61 for FA and KPCA (dimensions 92 and 112,respectively) and increases to 0.65 (dimension 647) for PCA. The F1-score increases from 0.66 to 0.70 for FA, 0.72 for PCA and 0.71 for KPCA at dimensions 92, 130 and 352, respectively. These F1-score increases can be mainly attributed to the large improvements of the recall.

Dimensionality reduction has a positive effect on the top 1 ranking performance (see Table 3.3). In summary, the relative success to rank the best LRMSD cluster in the top 1 improved from 38% in the standard ERT classifier to 46% for all three models ERT+FA, ERT+PCA and ERT+KPCA. The top 5 success rate dropped from 69% (standard ERT) to 61% for ERT+FA and ERT+KPCA and was unchanged for ERT+PCA. No improvements in the top 10 success rate could be obtained for all three tested transformations. The success rate decreases from 92% to 77% , 77% and 85% for ERT+FA, ERT+PCA and ERT+KPCA, respectively.

Also, the average ranking performance decreased for all three feature space transformations. Here, a decrease from 4.6 to 5.2 , 6.4 and 6.5 was yield for ERT+FA, ERT+PCA and ERT+KPCA, respectively. This drop in ranking ability

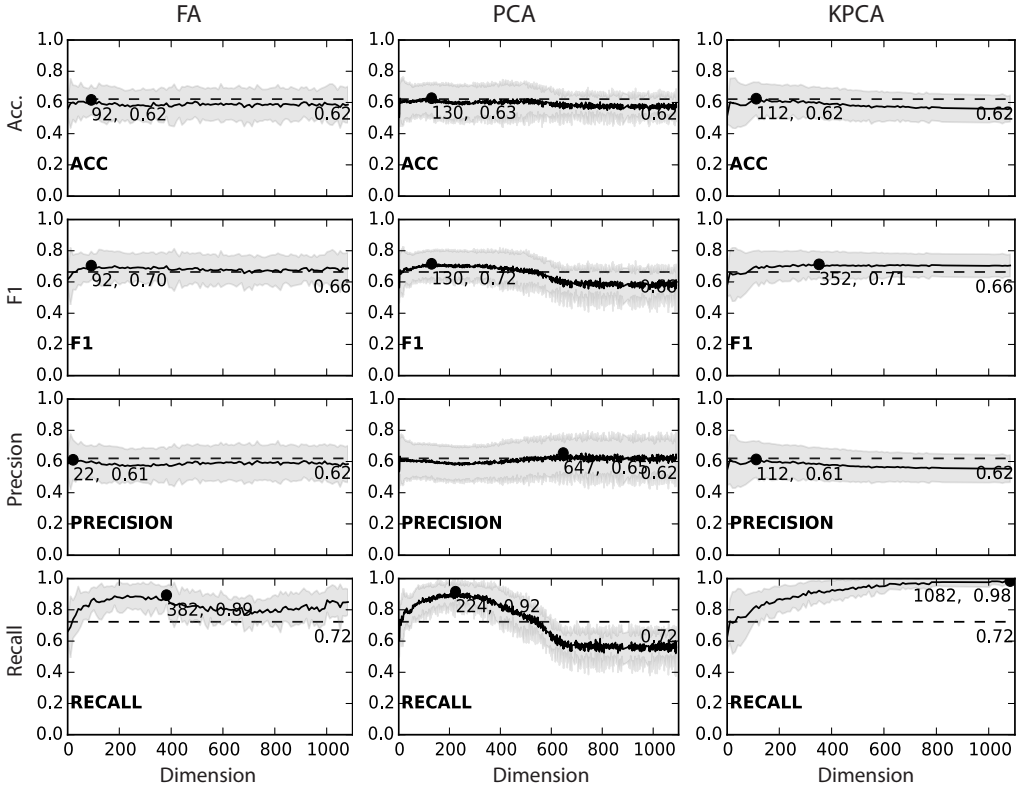

Figure 3.8: The change of accuracy, F1, precision and recall for feature space transformations with FA, PCA and KPCA with a rbf kernel. The solid black line shows the mean value and the light gray area indicates the standard deviation. The dotted gray line indicates the performance of the classifier without applied feature space transformation. Spheres indicate the best dimension with the highest value. Permission to reproduce this Figure has been granted by John Wiley & Sons, Inc.

can be attributed to the decrease in ranking accuracy for the two particularly hard targets T35 and T39. The rank for target T35 drops from 9 to 12 for all three feature space transformations. For T39, this change is even more pronounced, where the rank drops from 12 to 13 , 35 and 39 for ERT+FA, ERT+PCA and ERT+KPCA, respectively.

3.3.5 The Effect of Recursive Feature Elimination on Prediction Accuracy

Another method to reduce the dimensionality of the feature space is recursive feature elimination (RFE), where the features with the lowest feature importance

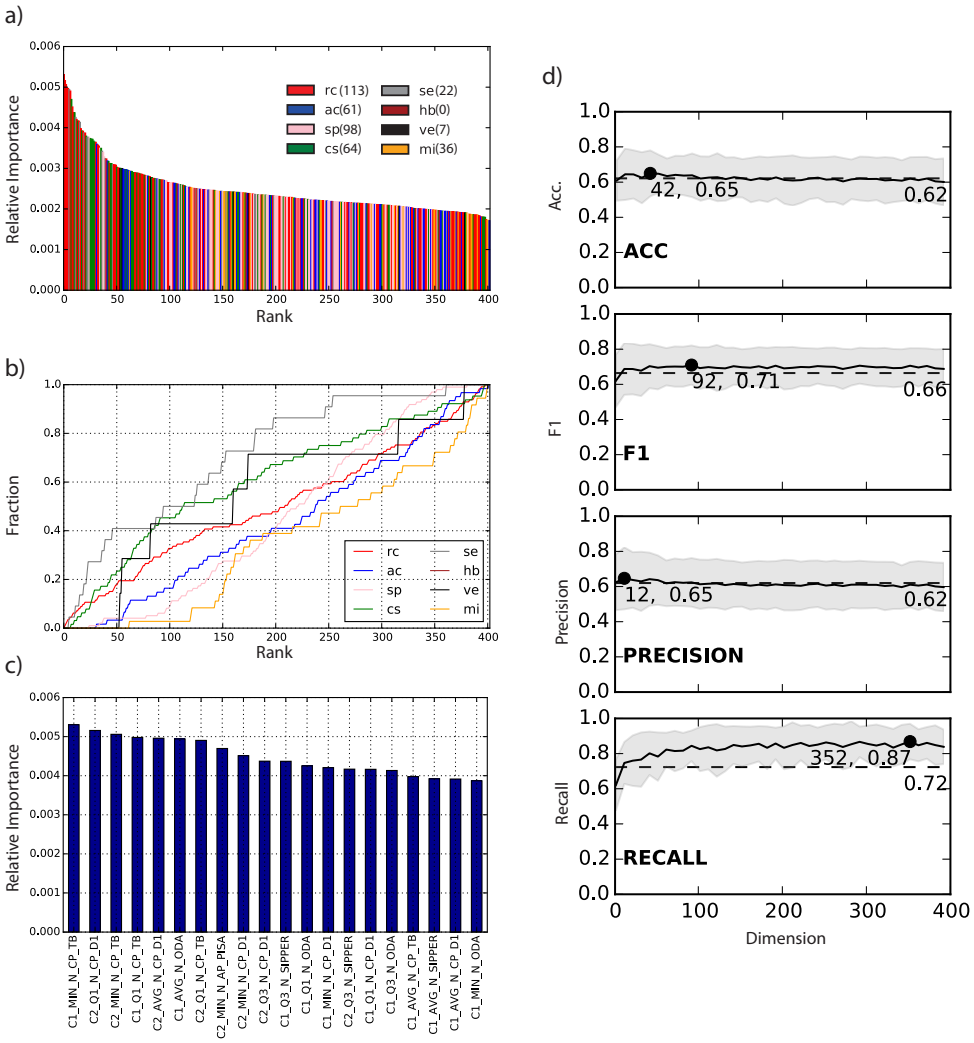

Figure 3.9: Analysis of the reduced feature set after RFE. (a) the relative feature importance for all 402 features colored by descriptor category. (b) fraction of features used from one of the 8 categories versus rank. (c) relative importance of the top 20 features. (d) change of accuracy, F1, precision and recall of the ERT+RFE+FA classifier for different dimensions. Permission to reproduce this Figure has been granted by John Wiley & Sons, Inc.

are iteratively removed from the set. Application of RFE to the training data yield the best F1-score at 402 dimensions. Training the ERT classifier on this reduced set of features, referred to as ERT+RFE, produced marked improvements in the top 1, top 10 and average ranking performance with respect to the standard ERT classifier. The top 1 ranking performance improved from 31% to 38% , for top 10 from 77% to 100% and the average rank improved from 4.6 to 3.9.

An analysis of the RFE shows that all features from category hb are removed and only 7 features from category ve remain in the set. In summary, the reduced set contains the following feature numbers when grouped by category: rc (113), ac (61), sp (98), cs (64), se (22), hb (0), ve (7) and mi (36). A feature ranking produced by sorting the reduced feature set in descending order according to feature importance shows that most of the top 50 features are associated with categories rc, se and cs (Figure 3.9a). This becomes visually more clear in the cumulative fraction versus rank plot in Figure 3.9b, where these three categories make up a high fraction early on. Features in the top 20 of the reduced set remain the same as in the full feature set were features from TOBI (6), DECK (7), ODA (4), PISA (1) and SIPPER (3) are present (Figure 3.9c).

Additionally, FA was applied to the reduced feature set in order to test whether the performance of the model can be further improved with feature space transformation (model ERT+RFE+FA). The results show that a marked improvement for recall was produced which had a maximum increase from 0.72 to 0.87 at dimension 352. For accuracy, F1 and precision the value improved from 0.62 to 0.65 (42 dimensions), 0.66 to 0.71 (92 dimensions) and 0.62 to 0.65 (12 dimensions), respectively. Rankings based on the ERT+RFA+FA model at 92 dimensions (i.e. best F1-score) yield improvements for both the top 1 and the top 5 success-rates, which improved from 38% to 46% and from 69% to 77%, respectively. However, the top 10 success rate decreased from 100% to 85%, which can be attributed to the stark increase in ranks for targets T35 and T39 where the best LRMSD cluster was ranked 12 and 33, respectively. Rankings for all other targets remained unchanged or improved as shown in Table 3.3.

3.4 Discussion

In this chapter, an integrated method to distinguish near native from incorrect docked poses clusters was presented. This method combines heuristic optimization with SwarmDock and predictive modelling with ERT. Localized SwarmDock enrichment was used to elevate the problem of power law distributions of cluster-size and to generate additional conformational poses to model the recognition process of protein-protein binding. Training of the ERT classifier was based on pair-wise cluster comparison where each cluster is described by 109 molecular descriptors. This pair-wise cluster comparison yield 7248 training examples. The overall results on the score_set dataset are promising. Compared to the scoring function DCOMPLEX, an improvement in relative ranking performance of 51% with the ERT+RFE was achieved.

3.4.1 Ranking with Statistical Learning

The focus of the presented method was to model the so called recognition process of protein-protein complex formation. Where the encounter complex samples a range of different conformations, rotations and translational poses in order to identify the true positive binding site. This is supported by the good performance of ensemble or cluster-based scoring schemes in previous CAPRI rounds (Oliva et al., 2013; Qin and Zhou, 2013). For example, ranking schemes that employed minimum cluster energies from a scoring function such as DCOMPLEX (i.e. used in SwarmDock) perform better than non-cluster based approaches.

Several issues are addressed which result from cluster-size imbalance and class bias of training examples that are a marked problem for many applications of machine-learning (Kubat and Matwin, 1997). The clustering by LRMSD applied to the decoy set of docked protein-protein complexes results in a power law distribution of cluster-size. Where a small number of clusters have a large number of models and most clusters have few models. This is visualized for each target in supplementary Figures B.1–B.13. This imbalanced distribution of

solutions can be an issue if conclusions are derived from distribution points such as median, first quartile, third quartile, minimum and maximum. In order to address this information bias, a localized enrichment of clusters with additional solutions was performed. Thus, resulting in more data-points describing the local energy landscape from the different molecular descriptors.

The second problem of class bias is defined as a large under representation of one class compared to another class during a supervised learning task. This results in a classifier with poor predictive sensitivity towards the under represented class when optimized for accuracy. Such a class-bias is present for protein-protein decoy sets such as the used score_set dataset where only 11% of all solutions are near-native. This issue is addressed by translating the learning task into a pairwise-comparison representation. In the presented method, every cluster n is compared to every other cluster m, resulting in an exhaustive set of unique comparisons. Class label assignments for the cluster comparisons are based on LRMSD, where the cluster with the smaller LRMSD is assigned the value 1 and the cluster with the higher LRMSD the value 0. This results in an almost equal distribution of class labels.

Overall, this learning representation has three advantages, namely i) this allowed the training of a classifier from a limited number of complexes, constructing a pair-wise comparison matrix generated 7248 training examples; ii) the class bias problem is resolved, which initially appeared from the small number of near-native versus incorrect clusters; iii) the method learns to rank implicitly according to LRMSD.

3.4.2 Physical Plausibility of the Model

The analysis presented in Section 3.3.3 shows that coarse grained potentials based on residue contacts such as the TOBI potential or the DECK scoring function have higher relative feature importance compared to fine grained potentials based on atomic contacts. Furthermore, functions describing the contribution of hydrogen bonding or VdW/electrostatic forces have limited predictive power. A possible

explanation for this can be the heterogeneity of the score_set dataset. The decoy ensemble for each target originates from a larger number of different protein-protein docking algorithms. Many of the models stem from rigid-body docking algorithms where usually an optimization in 6 dimensions is performed, i.e. translational and rotational, neglecting the possibility of conformational change. Hence, giving an advantage for coarse-grained functions. This is supported by Kuroda and Gray (2016) who performed a systemic analysis of docking accuracy as a function of backbone RMSD to bound conformation. The results showed that an RMSD of smaller than 0.6 Å is needed for classical energy functions to reliably identify the correct docking pose. Furthermore, it can be assumed that a large number of models in the score_set are not locally optimized by energy minimization or other refinement methods that would optimize for hydrogen bonding and VdW/electrostatic. Thus, making it hard to obtain reliable estimates for the identification of near-native and incorrect clusters.

3.4.3 Limitations

The method makes use of a cluster-size cutoff of ≤ 5. It was reasoned that energetically favourable patches on the receptor surface would produce more populated clusters instead of a single or a few solutions. However, there are also practical reasons for considering a cutoff value. Usually, the number of clusters for a single target can be several hundreds (see Table 3.1). Hence, a reduction in the number of clusters greatly reduces the associated computational cost of cluster enrichment and the computation of molecular descriptors. The disadvantage of such a strategy is that possible true-positive solutions could be removed. This, for example, happened for targets T35 and T39. Hence, a user of this methodology has to be aware of this limitation.

3.4.4 Future Optimizations

Parameters not systematically explored in this work are the cluster-size cutoff and the GROMOS parameter for cluster radius. The exploration of the first

parameter would help to estimate whether the predictors precision, recall and ranking performance would be positively or negatively affected when clusters < 5 are considered for training. The resulting additional cluster comparisons to train the ERT classifier could possibly help to build a model which has a higher precision at assigning the correct label for the pairwise-comparisons on the test targets. However, a negative effect can also be expected where the additional training data introduces noise by having too many irrelevant comparisons of incorrect vs. incorrect clusters, thus confusing the main learning objective of learning how to distinguish near-native binding sites from incorrect ones. The second parameter, with a possible impact on the classifiers performance, is the cluster radius. Currently, this parameter is 10 Å and was inspired by the LRMSD cutoff used for acceptable models in CAPRI. However, a systemic exploration of different cutoffs would help to clarify how important the local environment is for identifying the true positive binding site and which targets would benefit most. For example, an increase to 20 Å would allow for more extended clusters and larger deviations in receptor/ligand backbone conformations and potentially a more comprehensive description of the local energy landscape. Figure 3.5b indicates that complexes such as target T29 could possess a so ”called” funnel of attraction where surrounding clusters have energetic properties which rank them higher in cluster comparisons. Thus, exploiting this more systematically could greatly improve the success-rate.

Finally, improvements in the cluster enrichment process could also be beneficial. Currently, SwarmDock is employed to enrich initial cluster with additional solutions which differ in translation, rotation and conformation from the initial starting structure. Here, linear combinations of normal modes are used to generate new conformations. However, this approximation of flexibility might be too coarse grained to generate the transitions necessary to explore the local energy landscape in the detail required to make accurate predictions. A solution to this could be localized MD simulations that enrich the cluster in a physically more detailed and accurate way. An efficient way of doing so could be so called meta-dynamics simulations in contact map space as described in Chapter 5 of

this thesis, where a contact map definition of the binding interface is used to bias the simulation towards explorations of different contacts at the local binding site. However, extending this methodology to all clusters would require substantial computational and storage capacities.

CHAPTER 4

Optimization of Predicted Protein Folds by Refinement

4.1 Introduction

Protein structure prediction from sequence tries to overcome the limitations of experimental structure determination which are often time consuming and infeasible for certain types of proteins. Furthermore, construction of protein models seem to be the only practical solution for structural genomics where a high rate of newly discovered protein sequences demands for automated and fast structure determination (Baker and Sali, 2001). Current state of the art methods which make use of template based modelling (TBM) are partially successful as shown in several rounds of the Critical Assessment of Techniques for Protein Structure Prediction (CASP) competition (Moult et al., 2016; Huang et al., 2014; Moult et al., 2014; Mariani et al., 2011; Kryshtafovych et al., 2005; Moult, 2005). However, the quality of these TBM results are highly dependent on the presence of homologous proteins where at least one has been experimentially determined. An extension to TBM are so called refinement methods that further try to improve the initial models by extensively sampling new conformations. Essentially, emulating the later part of the protein folding pathway (see Figure 1.5). Physics based methods which make use of conformational sampling with molecular dynamics (MD) simulation have proven to

be successful in previous rounds of CASP (Feig, 2017; Modi and Dunbrack, 2016; Nugent et al., 2014). One such method by Mirjalili et al. (2014) makes use of restrained sampling with multiple replicated simulations and averages an ensemble of high scoring snapshots into one final refined model.

In this chapter, a sampling protocol based on molecular dynamics and meta-dynamics simulation is discussed. For CASP11 a scheme is explored which automatically derives restraints for the sampling from the set of submitted models in the TBM section. Here, structurally conserved regions in the set are identified and position restraints for residues or distance restraints between pairs of residues are derived. For CASP12, an enhanced sampling scheme in contact map space (CMS) is introduced, which is defined by observed intra residue-residue contacts and used as a collective variable (CV) to bias the potential of metadynamics simulations.

4.2 Methods

4.2.1 CASP11 Refinement Method

Essentially, the refinement method applied in CASP11 is based on restrained MD simulations with the goal to sample and decent the folding energy funnel. The restraints, which are based on position and residue-residue distance restraints, were automatically generated and are applied to structurally conserved regions that are identified from the set of given models for each refinement target. The flowchart in Figure 4.1a provides an overview of the method which can be explained as follows:

Download and filtering of all initial models For every refinement target all submitted TBMs from participating predictors in CASP11 were downloaded. The number of models varies from target to target, but on average 180 submissions were available. Often, a substantial part of these submissions were physically implausible, i.e. they contained long extended stretches. In order to avoid including these into the analysis a 10 Å Cα-RMSD cutoff to the provided starting model was applied.

Restraints generation Position and distance restraints for structurally conserved regions were generated and applied to Cα atoms. Position restraints were applied if the per-residue Cα RMSF calculated from the filtered set of TBM is < 3 Å. In order to determine conserved residue-residue distances all possible combinations of Cα-Cα pairs were measured and distance restraints were applied if all of the following criteria are true: a) the Cα-Cα pairs are at least 5 residues apart; b) the Cα-Cα distance is below 9 Å; c) the standard deviation of the distance is below 1 Å.

Sampling For each target three different simulation setups are executed: a) 3 ns long MD run without restraints, replicated 8 times; b) 3 ns long MD run with position restraints, replicated 8 times; c) 3 ns long MD run with distance restraints, replicated 8 times. All MD simulations were computed with GROMACS, using version 4.6 (Hess et al., 2008), and the G54a7 force field (Schmid et al., 2011). For all initial target structures hydrogen atoms were added and the systems were neutralized with Na^+ and Cl^- counter ions. A cubic simulation system with a 12 Å buffer between the edge of the box and the protein was solvated with TIP3P water molecules (Jorgensen et al., 1983). All targets were then subject to an energy minimisation using the steepest decent algorithm with a maximum of 50000 steps. This was followed by an equilibrium phase to relax the structure and its solvent. MD simulations (a) and (b) were subject to a 2 step equilibrium protocol where all heavy atoms were position restrained by a force of 1000 kJ $mol^{-1}nm^{-1}$ throughout the equilibration. In the first phase an NVT equilibration of the system was performed to increase the temperature from 0 K to 300 K in 100 ps using V-rescale (Bussi et al., 2007) for temperature coupling. The second phase consisted of a 300 ps long NPT equilibration of the system's pressure to 1 bar using Parrinello Rahman pressure coupling (Berendsen et al., 1984). For MD simulation (c) a second NPT equilibration was applied, where the first step consisted of a 200 ps long equilibration with full heavy atoms position restraints and distance restraints, and the second step of a 200 ps equilibration

with distance restraints only.

For all simulation setups a leap-frog integrator with a Δt of 2 fs was used and coordinates, velocities, energies, and forces were saved every 2 ps. Long range electrostatic interactions were treated with the Particle Mesh Ewald method (Darden et al., 1993) with a cutoff of 10 Å. Temperature and pressure coupling were controlled by the V-rescale and the Parrinello-Rahman method and were set to 300 K and 1 bar, respectively.

Snapshot selection Snapshots from each MD run were taken every 2 ps and scored with DFIRE (Liu et al., 2004). For each of the three different types of runs the 10 percent best snapshots were selected.

Model building Finally, five models were generated for every target: (a) an average model using the snapshots from the MD runs with distance restraints (abbreviated as ADR), (b) an average model using the snapshots from the MD runs with position restraints (abbreviated as APR), (c) an average model using the snapshots from the MD runs without restraints (abbreviated as ANR), (d) a model using the three average models as templates for the automated modelling with the software MODELLER (Eswar et al., 2008) (abbreviated as M3C) and (e) a model using the 5 best structures from the distance restrained MD runs (abbreviated as MDR). For the average models (a)-(c) the best 10 percent scoring structures of each MD run were used to calculate the average position for each atom. In order to resolve non-physical conformations from this averaging, a steepest decent energy minimization with a maximum of 50000 steps was applied.

4.2.2 CASP12 Refinement Method

The method tested in CASP12 made use of the structural variation present in the set of submitted predictions to infer restrains for conserved regions with low variability and to construct a contact map space (CMS) of observed residue-residue contacts in folds. This CMS is used as a collective variable (CV) in a metadynamic simulation

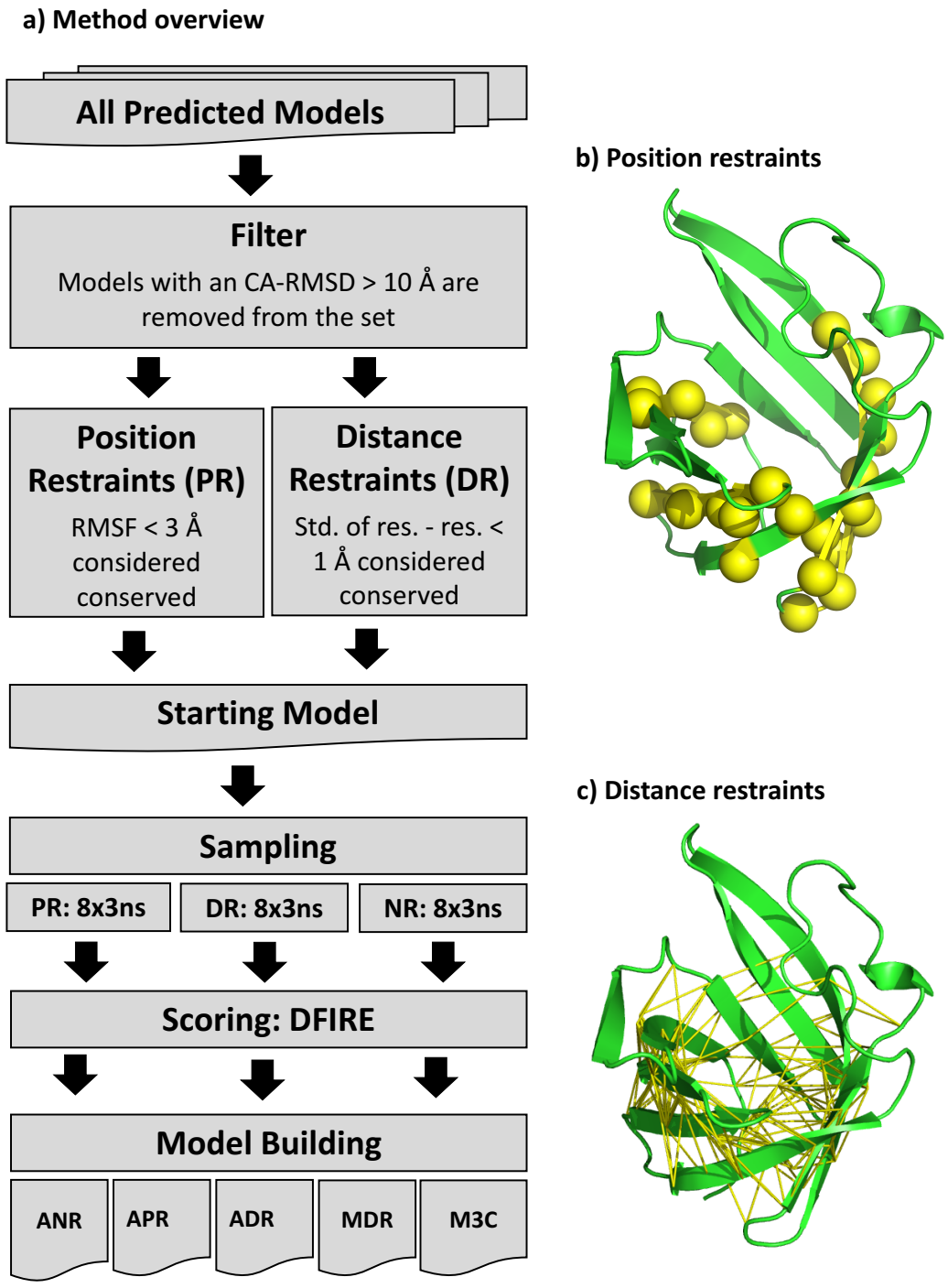

Figure 4.1: CASP11 method overview. a) Flowchart of the CASP11 method. b) Example of distance restraints for conserved residue-residue distances for target TR782, the yellow lines indicate the derived distance restraints for residue pairs. c) Example of position restraints for structurally conserved residue positions for target TR782. A detailed description of the method is provided in Section 4.2.1.

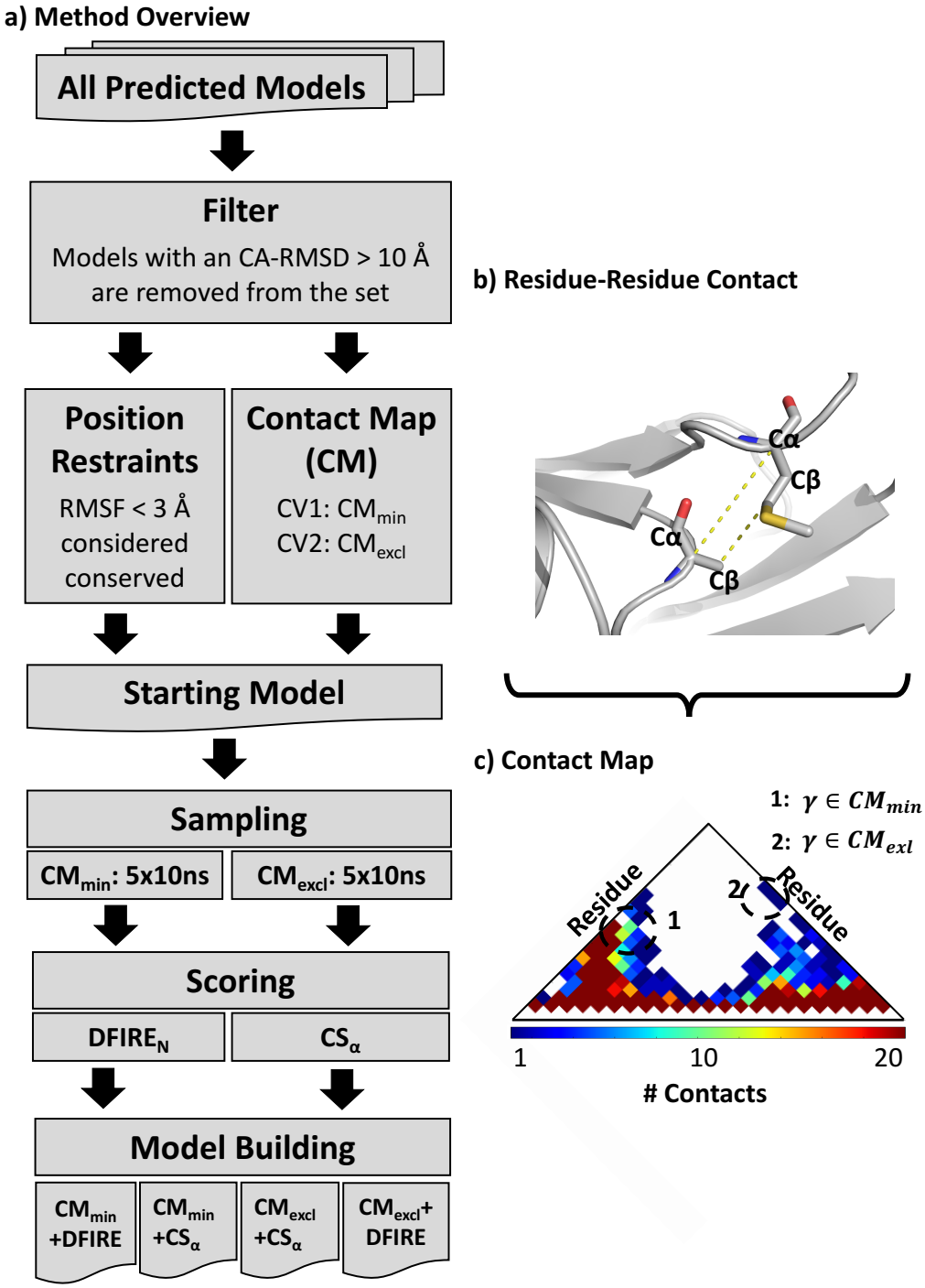

Figure 4.2: CASP12 method overview. a) Flowchart of the CASP12 method. b) Definition of intra residue-residue contacts, contacts considered between residues are based on Cα or Cβ distances below 8 Å, shown as dotted yellow lines. c) definition of contact map sets CM_{excl} and CM_{min} where the number of contacts for a contact $\gamma = 1$ and $\gamma \geq 1$, respectively. For CM_{min} if $\gamma > 1$ than the contact with the lowest distance is chosen as the reference r_{γ}^{0}. d) definition of the contact map space (CMS). A detailed description of the method is provided in Section 4.2.2.

for enhanced sampling and to reconstruct the conformational free energy landscape to guide model selection. This unique approach allows enhanced sampling around the variable of interest (i.e. reducing computational cost) and makes use of the information available from the different conformational states to guide the search for new energy minima. Essentially, the method can be divided into five parts that can be described as follows:

Filtering of all available models All available models from participating predictors of a target are downloaded from the prediction center server. Each model is compared to the starting model and the Cα RMSD is calculated, models with an RMSD > 10 Å are removed from the set.

Deriving position restraints The filtered set is used to determine structurally conserved residues. These are identified by computing the per residue root mean square fluctuation (RMSF) of Cα atoms. Residues with a RMSF < 3 Å are considered conserved and movements are restraint during the sampling process.

Contact map generation and collective variable definition From the structures in the filtered set of CASP predictions, residue-residue contacts are identified with a Cα or Cβ distance below 8 Å with the exception of direct neighbours, which are removed from the list. From these contacts two contact maps (CM) are generated, namely $\mathrm{CM_{exl}}$ and $\mathrm{CM_{min}}$. $\mathrm{CM_{exl}}$ contains contacts that are exclusive to one model from the filtered set, i.e. the contact is unique with a contact count of 1 (see Figure 4.2c). Whereas the map $\mathrm{CM_{min}}$ contains contacts with the lowest Cα/Cβ distance. From these CMs we can define two CVs describing the CMS:

$$CV1(R) = 1/N \sum_{\gamma \in CM_{exl}} (D_\gamma(R) - D_\gamma(R_{ref}))^2 \tag{4.1}$$

$$CV2(R) = 1/N \sum_{\gamma \in CM_{min}} (D_\gamma(R) - D_\gamma(R_{ref}))^2 \tag{4.2}$$

$$D_\gamma(R) = \frac{1-(r_\gamma/r_\gamma^0)^n}{1-(r_\gamma/r_\gamma^0)^m} \tag{4.3}$$

The sigmoid distance function $D_\gamma(R)$ is used to quantify the formation of a contact γ in structure R, where r_γ is the contact distance in structure R and r_γ^0 is the contact distance in reference structure R_{ref} which denotes to one of the models from the filtered set of CASP12 models where the contact was observed. Variables n and m are constant and set to $n = 6$ and $m = 10$.

Energy minimization, equilibration and sampling The preparation of the starting model prior to the sampling process follows a GROMACS standard procedure where the system is solvated, energy minimized and equilibrated for 300 ps. The sampling with metadynamics in CMS is performed at 300 K for 10 ns with 5 replicas for each CM definition, resulting in 100 ns sampling data for each target. The sampling of the CMS was performed with the GROMACS plug-in PLUMED2 (Tribello et al., 2014) where a Gaussian addition is deposited every 2 ps with $\sigma = 0.5$, a bias factor of 10 and an initial height of 5 kJ/mol.

Scoring and model building Snapshots from the trajectories are taken every 10 ps, resulting in 9810 frames in total. Scoring of these frames is based on reconstructing the free energy surface (FES) by integrating the deposited bias during the simulation and by computing the DFIRE energy. Furthermore, the combined scoring-function CS_α, that uses both normalized energies from FES and DFIRE to score frames, was used. Where $CS_\alpha = (1-\alpha)FES_N + \alpha DFIRE_N$ with an $\alpha = 0.5$ resulting in equal contribution of both terms for scoring.

4.2.3 Computation of GDTHA and RMSD

Model quality was assessed by computing the GDTHA and Cα-RMSD for all build models and snapshots of the trajectory. Details of these two metrics are explained in method Section 2.6.1. The reference crystal structure, starting model, snapshots

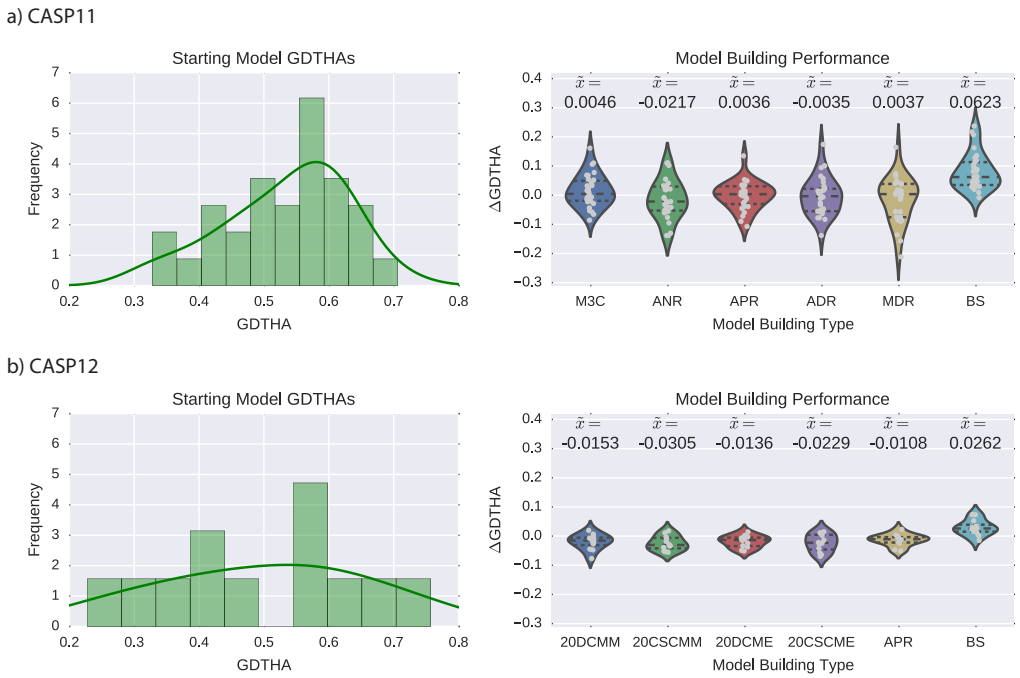

Figure 4.3: Model performance in CASP11 and CASP12. a) histogram on the left shows the starting model GDTHA for CASP11 targets and piano plot on the right shows the ΔGDTHA change after one of the 5 tested methods was applied. Additionally, the theoretical best change when selecting the best snapshot (BS) is also shown. b) histogram on the left shows the GDTHA of the starting models in CASP12. Piano plot on the right shows the ΔGDTHA change after one of the 5 methods was applied to the starting model and the theoretical best improvement when the best snapshot would have been selected (BS). The $\tilde{x}$ above each piano plot refers to the median value.

and final models were all stratified before metric computation. This stratification included the removal of atoms and residues which are not shared by all models in a target.

4.3 Results

4.3.1 Overall CASP11 Performance

The results shown in in Table 4.3 provide a complete overview of the CASP11 refinement targets for which a reference crystal structure was available. In total, an analysis for 30 CASP11 targets could be performed. The remaining 5 targets

Table 4.1: Restraints for CASP11 models. Shown are the target number (Target), number of residues (# of Res.), number of point restrained residues (# of PR), RMSF threshold used to define point restraints (PR Thresh.), number of distance restrained residues (# of DR), used distance retrained thresholds (DR Thresh.).

Target	# of Res.	# of PR	PR Thresh.	# of DR	DR Thresh.
TR217	224	154	0.2	138	1
TR228	84	10	0.4	19	2
TR283	168	20	0.3	8	1
TR759	62	0	0.3	2	1
TR760	210	137	0.3	131	1
TR762	257	219	0.3	174	1
TR765	76	12	0.3	6	1
TR768	143	79	0.3	84	1
TR769	97	-	-	0	1
TR774	155	58	0.3	25	1
TR776	219	169	0.3	133	1
TR780	95	74	0.3	56	1
TR782	110	26	0.3	0	3
TR783	243	155	0.3	87	1
TR786	217	108	0.3	87	1
TR792	80	43	0.3	14	1
TR795	136	111	0.3	93	1
TR803	134	4	0.3	6	1
TR810	243	182	0.4	157	1
TR816	68	32	0.4	2	1
TR817	489	356	0.3	352	1
TR821	255	52	0.3	160	1
TR828	84	9	0.4	4	1
TR829	67	39	0.5	18	1
TR833	108	56	0.3	59	1
TR837	121	11	0.5	0	1
TR848	138	58	0.3	68	1
TR854	70	56	0.3	10	1
TR856	159	92	0.3	58	1
TR857	96	20	0.3	0	1

Table 4.2: Restraints and CM for CASP12 models. Shown are columns for target number (Target), number of residue (# of Res.), number of point restrained residues (# of PR), the used point restrained threshold (PR Thresh.), number of residues in the CM_{excl} (# of Res. in CM_{excl}), number of contacts in the CM_{excl} (# of γ in CM_{excl}), number of residues in the CM_{min} (# of Res. in CM_{min}), number of contacts in the CM_{min} (# of γ in CM_{min}).

Target	# of Res.	# of PR	PR Thresh.	# of Res. in CM_{excl}	# of γ in CM_{excl}	# of Res. in CM_{min}	# of γ in CM_{min}
TR862	101	25	0.3	99	417	101	1149
TR868	116	16	0.3	115	493	116	1774
TR869	104	17	0.3	104	546	104	946
TR870	123	6	0.3	123	551	123	1860
TR872	88	37	0.3	85	267	88	1038
TR879	220	158	0.3	161	613	180	2186
TR891	119	98	0.3	67	93	101	327
TR893	169	123	0.3	149	371	163	1391
TR921	138	119	0.3	65	173	88	470
TR928	381	164	0.3	360	2149	376	7628
TR944	270	119	0.3	266	1543	270	4447
TR945	396	213	0.3	288	1165	329	3863

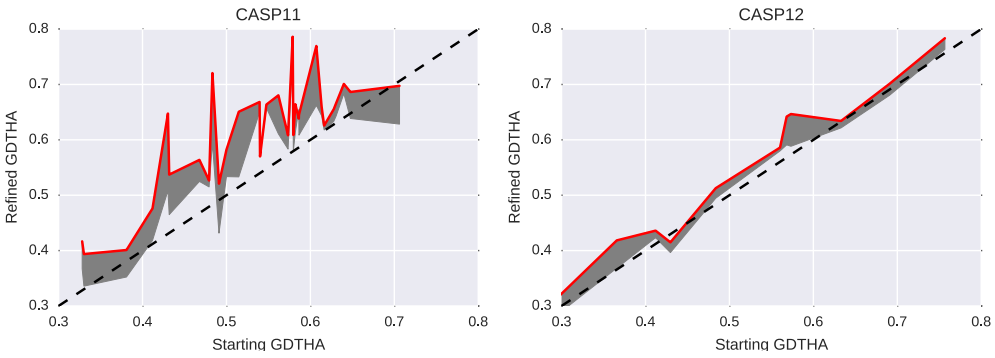

Figure 4.4: Starting GDTHA versus refined GDTHA. Plot on the left shows starting GDTHA (x-axis) versus refined GDTHA (y-axis) for CASP11 targets. Plot on the right shows the same but for CASP12 targets. The red line in both plots shows the theoretical best refinement improvement if the best snapshot would have been selected. The gray area indicates the actual refinement success from the best model generated by one of the 5 methods.

Table 4.3: GDTHA values for CASP11. Shown are the GDTHA for the starting model (SM), followed by the change after one of the 5 methods was applied (ADR, ANR, APR, M3C, MDR). Furthermore, the theoretical best improvement if the best snapshot would have been selected is shown (Best) and its source sampling method (Source). Bolt numbers indicate an improvement over the SM GDTHA. The reported success-rate for each model building strategy is defined as the number of times an GDTHA improved divided by the total number of targets. A higher value indicates better performance. The rank of the best snapshot is shown in supplemental material Table C.1.

Target	SM	ADR	ANR	APR	M3C	MDR	Best	Source
TR217	0.628	0.576	0.573	**0.634**	0.613	0.492	**0.656**	PR
TR228	0.548	**0.613**	**0.658**	**0.601**	**0.616**	**0.586**	**0.664**	NR
TR283	0.412	**0.418**	0.372	0.322	0.405	0.402	**0.476**	DR
TR759	0.430	**0.475**	**0.471**	**0.480**	**0.508**	**0.504**	**0.648**	PR
TR760	0.573	0.435	0.443	**0.583**	0.559	0.362	**0.608**	PR
TR762	0.706	0.628	0.568	0.599	0.621	0.550	0.698	PR
TR765	0.579	**0.753**	**0.681**	**0.714**	**0.740**	**0.743**	**0.786**	NR
TR768	0.640	**0.657**	**0.670**	0.568	**0.682**	**0.684**	**0.701**	PR
TR769	0.562	0.552	0.554	**-**	0.559	**0.611**	**0.680**	NR
TR774	0.381	0.334	0.321	0.336	0.352	0.305	**0.401**	PR
TR776	0.631	0.576	0.578	0.603	0.587	**0.643**	**0.666**	PR
TR780	0.540	**0.634**	0.518	0.516	**0.650**	**0.605**	**0.668**	DR
TR782	0.648	**-**	0.596	0.609	0.639	**-**	**0.686**	PR
TR783	0.586	**0.617**	**0.607**	**0.621**	**0.654**	**0.596**	**0.638**	PR
TR786	0.479	0.396	0.403	**0.515**	**0.489**	0.393	**0.527**	PR
TR792	0.607	0.601	**0.662**	0.568	**0.659**	**0.623**	**0.770**	DR
TR795	0.586	**0.609**	**-**	0.574	**0.609**	**0.590**	**0.645**	DR
TR803	0.330	**0.336**	0.299	**0.336**	0.310	**0.332**	**0.394**	DR
TR810	0.540	**0.553**	**0.559**	**0.554**	**0.573**	0.483	**0.570**	PR
TR816	0.515	0.489	0.489	0.507	0.500	**0.533**	**0.651**	DR
TR817	0.468	**0.525**	**0.487**	**0.469**	**0.522**	**0.477**	**0.564**	DR
TR821	0.483	**0.584**	**0.594**	**-**	**0.589**	**0.525**	**0.721**	NR
TR828	0.491	0.414	0.396	0.432	0.426	0.402	**0.521**	PR
TR829	0.500	0.466	0.463	**0.534**	0.478	**0.504**	**0.582**	PR
TR833	0.613	0.556	0.572	**0.644**	0.556	0.537	**0.660**	PR
TR837	0.432	**0.446**	**0.465**	**0.446**	**0.453**	**0.444**	**0.537**	DR
TR848	0.580	0.516	0.558	0.578	0.580	0.524	**0.609**	PR
TR854	0.582	0.579	**0.614**	**0.614**	**0.614**	0.568	**0.664**	NR
TR856	0.616	0.561	0.478	**0.618**	0.563	0.531	**0.626**	PR
TR857	0.328	**0.344**	0.318	**0.341**	**0.344**	**0.367**	**0.417**	PR
Success-rate		0.467	0.367	0.533	0.500	0.567	0.967	
Median (all)		-0.003	-0.022	0.004	0.005	0.004	0.062	
Median (imp.)		0.026	0.033	0.022	0.047	0.018	0.064	

TR274, TR280, TR811, TR823 and TR827 are excluded. These 30 targets have a wide range of starting GDTHA ranging from 0.328 to 0.706 with most starting models in the range of 0.55 to 0.6 (see histogram in Figure 4.3a). Overall, 25 out of 30 targets could be improved in at least one of the 5 build models. However, the success rate and median improvement of GDTHA varies substantially between the different model building strategies. The most successful strategy is MDR with a relative success rate of 0.567 followed by APR (0.533), M3C (0.5), ADR (0.467) and ANR (0.367). The piano plot in Figure 4.3a shows that the median improvement of GDTHA is only positive for 3 out of 5 strategies. Where M3C could yield the largest median improvement of 0.005 GDTHA points. This is followed by APR and MDR with a median improvement of 0.004 GDTHA points each. However, the two methods ANR and ADR were not able to yield a positive median model improvement and resulted in a -0.022 and -0.003 GDTHA decrease, respectively. When only successful refinements are considered for the calculation of the median the best method M3C yields an improvement of 0.047 GDTHA points followed by ANR (0.033), ADR (0.026), APR (0.022) and MDR (0.018).

The theoretical best success-rate, if the best snapshot as generated by the different MD based sampling approaches would have been selected, yields a success rate of 0.967 where only one target (TR762) was not improved. The theoretical best median GDTHA improvement is 0.062 points for all targets and 0.064 points when only success-full refinement targets are considered. This shows that the sampling via MD is able to generate markedly improved conformations. The source of these best GDTHA snapshots are with a large majority from position restrained (PR) MD simulations which generate the best snapshot 17 times, followed by distance restrained (DR) simulations with 8 and non-restrained (NR) simulations with 5.

The refinement method as applied in CASP11 is able to improve the model quality for a wide range of starting GDTHA (see Figure 4.4). Though, models with very low starting quality in the range from 0.3 - 0.4 GDTHA show less improvement. A similar drop in refinement success is observed for high quality models in the range of 0.65 - 0.75 GDTHA which improved only slightly, or even

decreased in quality after refinement.

4.3.2 Overall CASP12 Performance

Table 4.4: GDTHA values for CASP12. Shown are the GDTHA for the starting model (SM), followed by the change after one of the 5 methods was applied (20CSCME, 20CSCMM, 20DCME, 20DCMM, APR). Furthermore, the theoretical best improvement if the best snapshot would have been selected is shown (Best) and its source sampling method (Source). Bold numbers indicate an improvement over the SM GDTHA. The reported success-rate for each model building strategy is defined as the number of times an GDTHA improved divided by the total number of targets. A higher value indicates better performance. The rank of the best snapshot is shown in supplemental material Table C.1.

TR	SM	20CS CME	20CS CMM	20D CME	20D CMM	APR	Best	Source
TR862	0.366	**0.368**	0.363	0.363	0.350	0.363	**0.418**	CMM
TR868	0.573	**0.588**	0.535	0.567	0.573	0.569	**0.647**	CME
TR869	0.289	0.264	0.255	0.274	0.262	0.276	**0.305**	CME
TR870	0.228	0.199	0.195	0.203	0.183	0.224	**0.262**	PR
TR872	0.568	**0.580**	0.540	**0.582**	0.546	**0.591**	**0.642**	CMM
TR879	0.633	0.564	0.615	0.584	0.618	0.622	**0.634**	CMM
TR891	0.757	0.714	0.703	0.705	**0.763**	0.710	**0.784**	CME
TR893	0.691	0.627	0.639	0.654	0.614	0.681	**0.701**	CME
TR921	0.484	**0.495**	0.476	0.471	0.475	0.462	**0.513**	CMM
TR928	0.430	0.372	0.374	0.396	0.382	0.379	0.415	CME
TR944	0.560	0.539	**0.578**	**0.562**	**0.580**	0.537	**0.586**	PR
TR945	0.412	0.398	**0.423**	0.401	0.405	**0.415**	**0.436**	CMM
Success-rate		0.333	0.167	0.167	0.167	0.167	0.917	
Median (all)		-0.023	-0.030	-0.014	-0.015	-0.011	0.026	
Median (imp.)		0.011	0.015	0.008	0.010	0.013	0.027	

The refinement of CASP12 targets proofed to be harder than in previous CASP rounds. In this round the new sampling and model building strategies tested here were only able to yield an improved model in 7 out of the 12 targets which were available for analysis. Results for other targets could not be manually analysed because of missing reference X-ray structures. The range of starting GDTHAs for these targets was more diverse compared to CASP11. The set of refinement

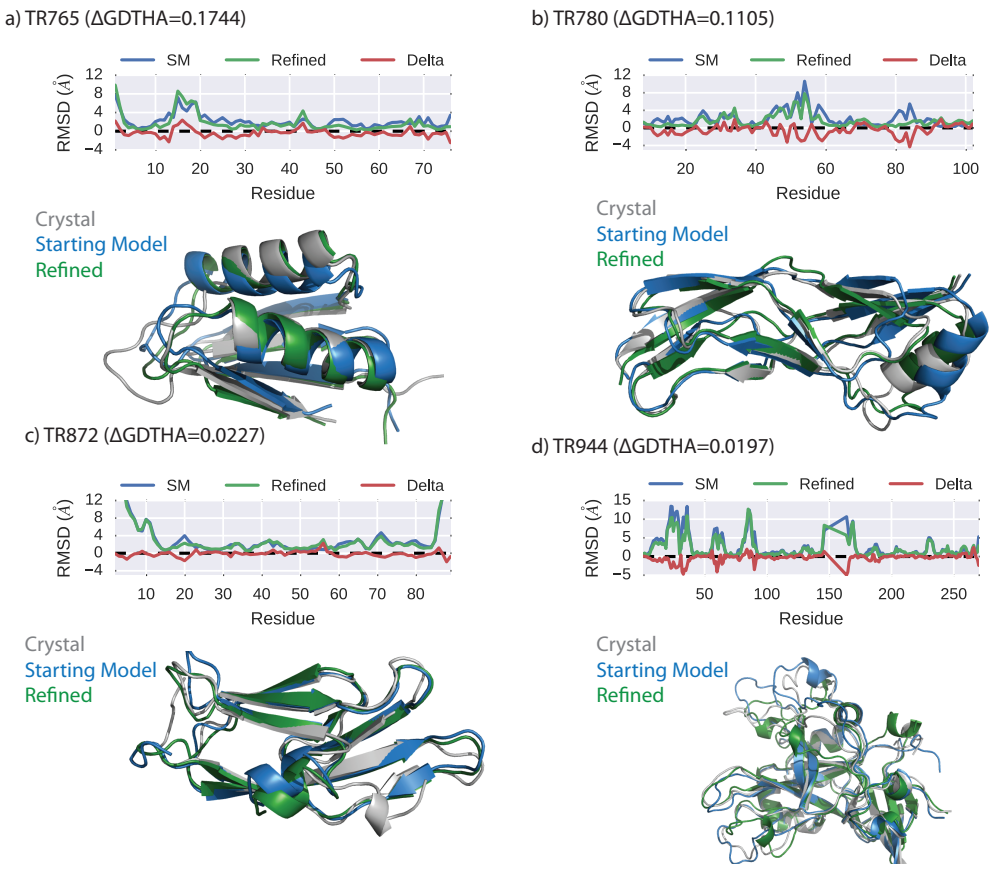

Figure 4.5: Refinement examples from CASP11 and CASP12. Shown are the refinement success for two CASP11 and CASP12 targets: a) TR765, b) TR780, c) TR872, d) TR944. Each sub-plot shows on top the per residue RMSD for SM in blue, refined model in green and the delta RMSD between SM and refined in red. The bottom part shows a superimposed a 3D rendering of the models from the experimental crystal structure in gray, the starting model in blue and the refined model in green.

targets included quasi random folds such as TR870 with a GDTHA of 0.228 and very high-quality models such as TR891 with a GDTHA of 0.757. A complete overview of all starting models and their refinement success is available in Table 4.4 and Figure 4.3b. Overall, the refinement success for each individual method was low. As a reference, the in CASP11 established refinement method APR was also included. The success-rate for this strategy dropped from 0.533 in CASP11 to 0.167 in CASP12 to produce a model with improved quality and shows the increased difficulty for refinement of these targets. The highest success-rate of 0.333 was achieved by the newly introduced strategy 20CSME which makes use of sampling in contact map space and the new scoring function CS_α. All other strategies

(20CSCMM, 20DCME, 20CMM, APR) had a success-rate of 0.167. As a result of the low success rate no model building strategy has a positive median ΔGDTHA. The best performing method is APR with -0.0108 median ΔGDTHA, followed by 20DCME (-0.0136), 20DCMM (-0.0153), 20CSCME (-0.0229) and 20CSCMM (-0.0305). When considering only successful refined targets for median calculation the best strategy is 20CSCMM with a median ΔGDTHA of 0.015 closely followed by APR (0.013), 20CSME (0.011), 20DCMM (0.010) and 20DCME (0.008).

The theoretical best success-rate for generating an improved model is 0.917 if the best snapshot would have been selected. Here, all but one target (TR928) improved. This theoretical best improvement results in a median ΔGDTHA of 0.026. Interestingly, the two new sampling methods based on sampling in CMS (i.e. CMM and CME) seemed to be most successful in producing new conformations with improved quality compared to the protocol based on position restraints (PR). Sampling based on CMM and CME produced 10 times the snapshot with the largest improvement whereas PR was only able to generate this snapshot for 2 targets.

The analysis of starting GDTHA versus refined GDTHA shown in Figure 4.4 shows that models with a wide range of GDTHAs could be improved. However, the extent of improvement was significantly lower compared to CASP11.

4.3.3 Secondary Structure and Amino Acid Dependency for Successful Refinement

Refinement in CASP11 is successful for all three secondary structure elements helical (H), β-strand (E), and coil (C). The heatmap in Figure 4.6b shows the average ΔRMSD based on the best GDTHA snapshot generated for each target. The bins group the results into different initial deviations as measured in the starting model. The ΔRMSD improves for all bins but 0-1 Å where a small decrease in ΔRMSD was measured with values of 0.21 Å, 0.06 Å and 0.07 Å for C, E and H, respectively. The largest improvement was observed for E in the 4-5 Å bin. A more detailed look at the refinement success for individual amino acid types shows that most of the amino acids can be improved for all bins but 0-1 Å (see Figure 4.6).

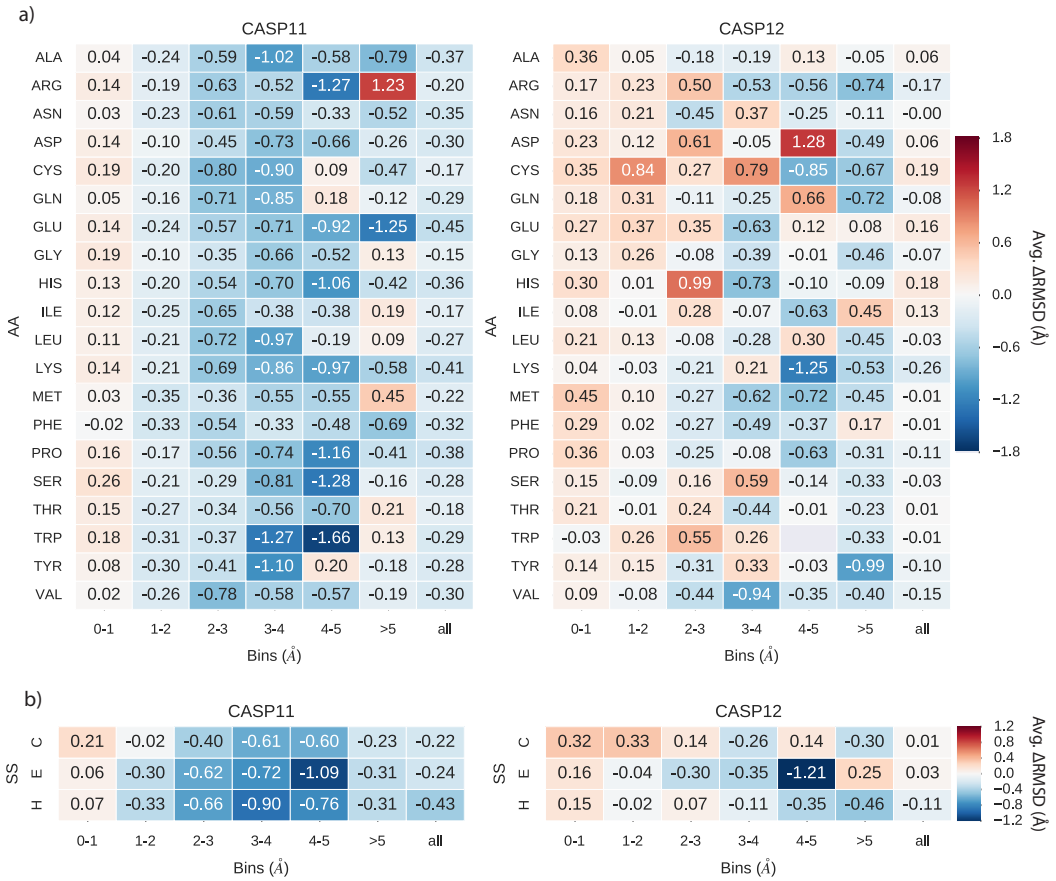

Figure 4.6: Refinement success as a function of secondary structure and amino acid composition. a) shows a heatmap of average RMSD refinement success for CASP11 (left) and CASP12 (right) targets of all 20 amino acids and different bins of deviation as observed in the starting model. b) shows a heatmap of average RMSD refinement success for different secondary structure elements and bins for CASP11 (left) and CASP12 (right). The secondary structure assignment is based on DSSP (Kabsch and Sander, 1983)

The best improvement was yield by bin 4-5 Å for amino acid tryptophan with an average ΔRMSD of -1.66. However, a markedly negative refinement success could be observed for arginine in bin >5 Å where an increase of 1.23 Å was measured.

Results for CASP12 show less refinement success for optimizing different secondary structure elements. The best improvement yield refinements of helices where an overall average improvement of −0.11 Å ΔRMSD is measured. Refinement is also successful for most β-sheet bins. Improvements of −0.04 Å, −0.30 Å, −0.35 Å and −1.21 Å in ΔRMSD for bins 1-2 Å, 2-3 Å, 3-4 Å, 4-5 Å are observed, respectively. Less successful was the refinement of loop regions.

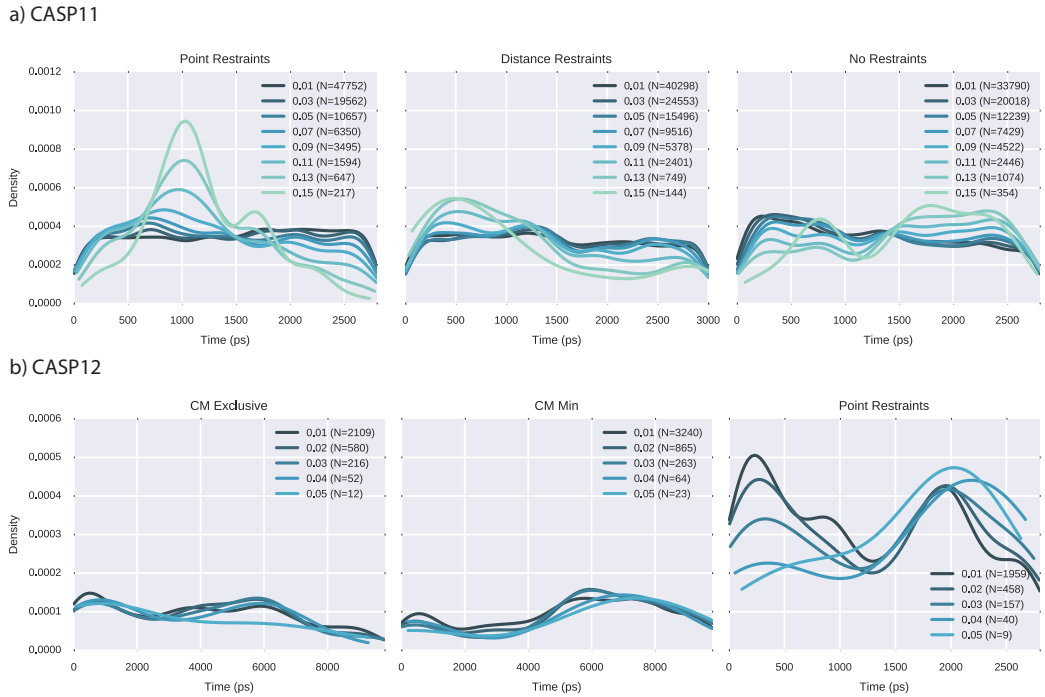

Figure 4.7: Refinement success versus time. Shown are density plots for different sampling methods tested in a) CASP11 and b) CASP12. The different coloured lines indicate the density of GDTHA improvement at different threshold levels. For example, the line indicating 0.01 considers all snapshots with a ΔGDTHA > 0.01. The number in brackets indicates the number of snapshots in this category.

Here, only bins 3-4 Å and >5 Å showed improvements of -0.26 Å and -0.30 Å, respectively. Similar mixed results are shown in the heatmap for different amino acids types in Figure 4.6a in CASP12. For bins 0-1 Å and 1-2 Å the ΔRMSD decreased for most types. However, the results become better for bins with higher initial deviations of the starting model from the reference structure. For example, most amino-acid types are improved for bins 4-5 Å and >5 Å.

4.3.4 How Much Sampling is Needed for Successful Refinement?

The results shown in the histogram in Figure 4.7 visualize the density of snapshots with ΔGDTHA improvements ranging from 0.01 to 0.15 in increments of 0.02 for CASP11 and from 0.01 to 0.05 in increments of 0.01 for CASP12. The decreased range of values for CASP12 is due to the less successful sampling of snapshots with

larger improvements.

In CASP11 sampling based on NR produced the highest number of snapshots with ≥ 0.15 ΔGDTHA where N=354. The two other sampling methods PR and DR were able to generate N=217 and N=144 poses in this category. Interestingly, sampling based on PR has a strong tendency to generate these high-quality conformations in the 700 ps - 1500 ps time range peaking at 1000 ps in their 8 times replicated 3000 ps long trajectories. Likewise, sampling from DR trajectories also has a strong propensity to produce snapshots at earlier time-points (200ps-1000 ps). Showing that long simulation times are not necessary for sampling with DR and PR to produce conformations with large ΔGDTHA improvements. Contrary to this observation is unrestrained sampling (NR). Here, snapshots with the highest improvement have the largest propensity at the 1800 ps-2500 ps time range. Snapshots with smaller improvements (0.01-0.05) are equally distributed in all three MD based sampling methods tested in CASP11.

Sampling of improved snapshots was less successful in CASP12, indicating an improved difficulty of the targets selected for refinement. The quantity and magnitude of generating improved conformations is significantly smaller compared to the previous CASP round. For example, the sampling strategies tested here (CME, CMM, PR) were only able to generate snapshots with up to 0.05 ΔGDTHA improvements where N=32 for CMM, N=12 for CME and N=9 for PR. The sampling with CME and CMM produced a flat density profile for improved snapshots across the sampling time of 5 times replicated 10 ns MD simulations. For sampling with PR, better quality snapshots (0.04 and 0.05) are only sampled at the end of the simulation time.

4.3.5 Dependency of Refinement Success with Model Source

The refinement was successful for targets originating from a diverse set of methods in CASP11 (see Table 4.5). The refinement was successful for 11 out of 13 model sources. The largest average ΔGDTHA improvement with 0.110 was obtained for the SAM-T08-server (Karplus, 2009) which provided one model. This method is

Table 4.5: Average refinement success as a function of model source. The table shows the method name, number of models included in the analysed set for that method (Count), their average initial GDTHA and RMSD value and the average ΔGDTHA and ΔRMSD after refinement for the best model.

Method	Count	Avg. Initial GDTHA	Avg. Initial RMSD	Avg. ΔGDTHA	Avg. ΔRMSD
CASP11					
SAM-T08-server	1	0.548	3.918	0.110	-0.768
QUARK	4	0.513	3.594	0.088	-0.301
PhyreX	1	0.586	3.073	0.068	-0.117
myprotein-me	3	0.474	3.415	0.057	-0.284
nns	2	0.421	4.097	0.042	-0.442
eThread	1	0.328	4.061	0.039	-0.085
Atome2_CBS	2	0.548	2.571	0.031	-0.339
MULTICOM-CLUSTER	1	0.613	4.715	0.030	-1.266
BAKER-ROSETTASERVER	6	0.577	4.478	0.017	-0.227
MULTICOM-NOVEL	2	0.580	2.762	0.016	-0.045
Zhang-Server	4	0.546	2.613	0.002	0.045
RaptorX	2	0.673	2.839	-0.017	-0.357
FALCON_MANUAL	1	0.381	4.968	-0.029	0.191
CASP12					
Pcons-net	1	0.568	5.589	0.023	-1.051
BAKER-ROSETTASERVER	1	0.573	3.010	0.015	-0.767
GOAL	6	0.532	5.348	0.004	-0.033
QUARK	1	0.366	5.921	0.003	-0.125
BhageerathH-Plus	1	0.228	9.430	-0.004	-0.430
FFAS-3D	1	0.633	5.503	-0.011	-0.456
HHPred0	1	0.430	5.965	-0.033	0.090

based on template based modelling and finds homologs based on a hidden Markov model generated from a multiple sequence alignment.

Methods for which no improvement was yield are RaptorX (Ma et al., 2013) and FALCON_MANUAL (Li et al., 2008) with an avg. ΔGDTHA of -0.017 and -0.029, respectively. RaptorX is a threading method that uses a context-specific alignment potential. The other method, FALCON_MANUAL, is a fragment assembly method that uses a position-specific hidden Markov model to predict the protein structure.

The 12 reference structures available for analysis in CASP12 originate from 7 different methods where GOAL (Joo et al., 2015) has the highest count with 6 and all other methods contributed 1 starting model each. Here, refinement was successful for 4 out of the 7 methods. The best refinement success was possible for Pcons-net with a ΔGDTHA of 0.023.

4.3.6 CASP Post-Mortem: Optimizing for the Number of Snapshots for Model Building

During CASP11 models for methods based on averaging (i.e. ADR, ANR and APR) used 10 percent of the best scoring snapshots. This value was inspired from Mirjalili et al. (2014) and not further empirically tested whether other values produce better results. Similarly, in CASP12 values of 10 percent were chosen for APR and 20 snapshots for methods 20CSCME, 20CSCMM, 20CME and 20DCMM. In order to investigate whether other number of snapshots produce on average better results for CASP11 and CASP12 targets a series of snapshots ranging from best 1 snapshot to best 20 percent of the trajectories total number of snapshot for model building were tested. Here, snapshots are ranked according to their DFIRE energy.

For CASP11 targets the best performance based on average ΔGDTHA is at 1770 snapshots (avaerage 0.0028 ΔGDTHA), 2350 snapshots (average -0.0074 ΔGDTHA), and 1190 snapshots (-0.0115 ΔGDTHA) for PR, DR and NR, respectively. For ΔRMSD based performance quantification the best number of snapshots are at 1030 (-0.0673 ΔRMSD), 2350 (-0.0690 ΔRMSD) and 2070

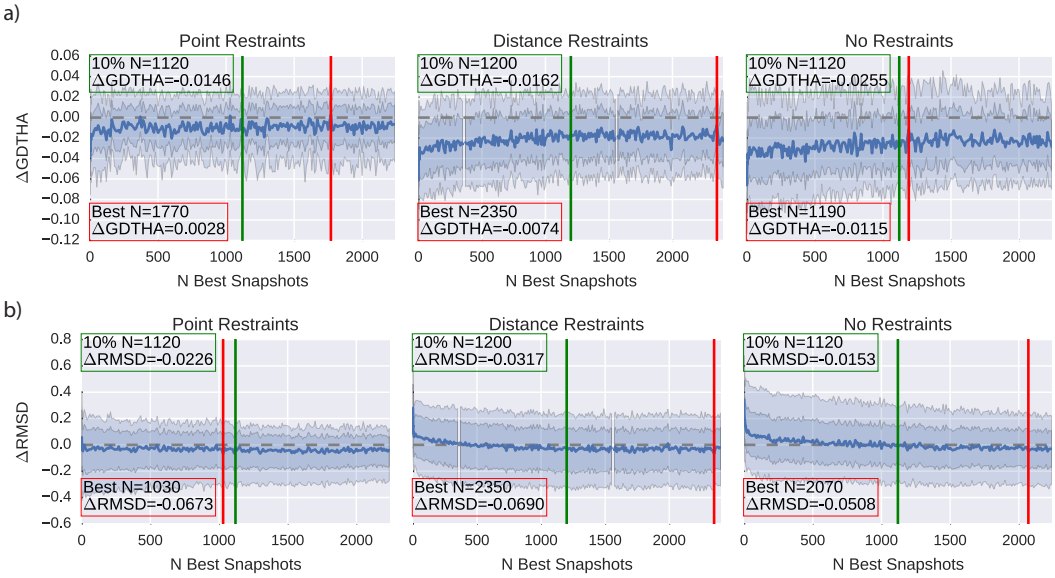

Figure 4.8: Average model quality as a function of selected snapshots (CASP11). a) average ΔGDTHA after model building using the N best ranked snapshots according to DFIRE. b) average ΔRMSD after model building using the N best ranked snapshots according to DFIRE. The two transparent bands in each sub-plot indicate the confidence interval at 95 and 100 percent. The green line and box indicate the number of snapshots used in CASP and the red line and box indicate the best number of snapshots.

(-0.0508 ΔRMSD) for PR, DR and NR, respectively.

In CASP12, the increase from the original 20 snapshots to a larger number has a remarkably positive effect on the average ΔGDTHA and ΔRMSD performance. The best performance based on ΔGDTHA are obtained at 2590, 3810 and 2120 snapshots for CME (-0.0064 ΔGDTHA), CMM (-0.0125 ΔGDTHA) and PR (-0.0114 ΔGDTHA), respectively. Similarly, best ΔRMSD values are yield at 2550 snapshots (-0.998 ΔRMSD), 4650 snapshots (0.1211 ΔRMSD) and 2120 snapshots (0.0867 ΔRMSD) for CME, CMM and PR, respectively.

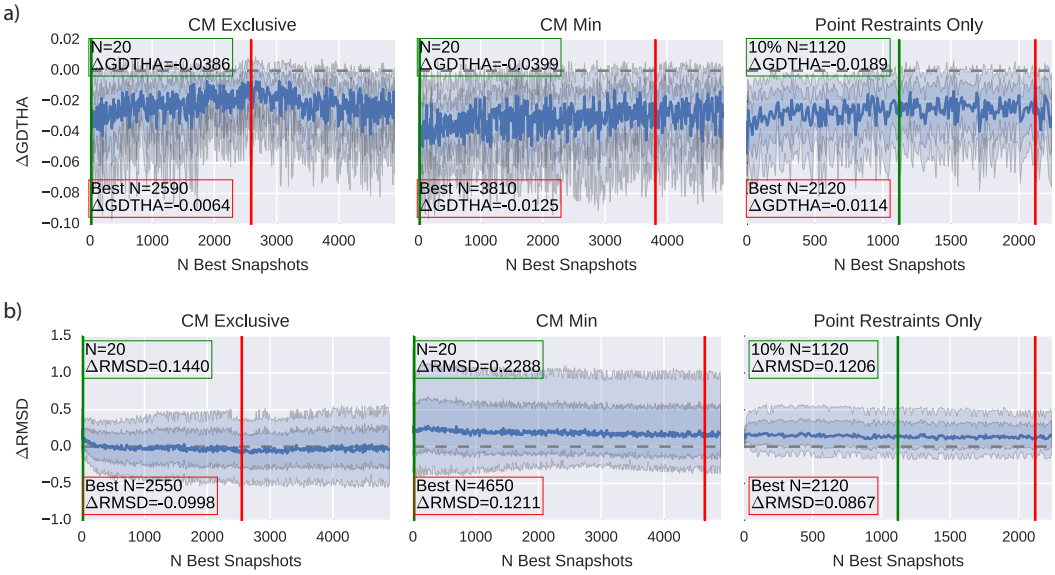

Figure 4.9: Average model quality as a function of selected snapshots (CASP12). a) average ΔGDTHA after model building using the N best ranked snapshots according to DFIRE. b) average ΔRMSD after model building using the N best ranked snapshots according to DFIRE. The two transparent bands in each sub-plot indicate the confidence interval at 95 and 100 percent. The green line and box indicate the number of snapshots used in CASP and the red line and box indicate the best number of snapshots.

4.4 Discussion

4.4.1 MD Based Sampling is Successful in Generating Improved Conformations

The results from CASP11 and CASP12 demonstrate the feasibility of MD based sampling to generate snapshots with improved quality for 96.7 percent of analysed CASP11 targets and 91.7 percent of analysed CASP12 targets. The three tested sampling methods in CASP11 based on different restraint types DR, PR and NR showed that PR has the highest rate among all analysed targets to generate the snapshot with the largest improvement (i.e. 17 out of 30, see Table 4.3). The two other sampling methods DR an NR have a significantly lower success-rate with 8 and 5 to generate the snapshot with the largest improvement. The unrestrained MD samplings (NR) cause often a drift of parts of the structure which are already in

good agreement with the crystal structure. To be precise, NR sampled N=33790 snapshots with a ΔGDTHA improvement of > 0.01. This is a marked lower number compared to restrained sampling methods, such as PR and DR which produced N=47752 and N=40298, respectively. However, when NR sampling is successful for a target it has a higher potential of generating a snapshot with large improvements. For example, NR generated a snapshot with a ΔGDTHA improvement of 0.207 for target TR765 where the model quality improved from 0.579 to 0.786 GDTHA. This is also reflected when looking at the overall number of snapshots with a ΔGDTHA > 0.15 for PR, DR and NR where N=217, N=144 and N=354, respectively.

In CASP12, two new enhanced sampling methods were tested, i.e. CME and CMM, that make use of sampling in a predefined CMS of observed intra residue-residue contacts. In theory, this should allow for more directed sampling and result in models with larger improvements compared to e.g. PR based sampling. However, during CASP12 all three sampling methods failed to perform equally well during refinement compared to CASP11. The number of sampled snapshots with improvement and the magnitude of improvement sampled decreased significantly. In the 12 targets available for analysis the number of snapshots with a ΔGDTHA > 0.01 was N=2109, N=3240 and N=1959 for CME, CMM and PR, respectively. This shows that CME and especially CMM were better at sampling improved snapshots compared to PR. Interestingly, these two sampling methods were more successful at sampling ΔGDTHA > 0.05 snapshots with N=12, N=23 and N=9, respectively. This provides partial evidence that these two sampling methods are better at producing snapshots with higher quality improvements. This is also shown in Table 4.4 where the best snapshot was generated 10 times by either CMM or CME for the 12 analysed targets. However, since the refinement success between CASP11 and CASP12 targets was significantly different it is hard to quantify how much better CMM and CME are compared to, e.g. PR.

4.4.2 Limitations of Energy Based Snapshot Selection

Reliable identification of improved snapshots with scoring or energy functions is challenging. Often the energy landscape as quantified, for example, by DFIRE is inaccurate where models with a better agreement to the reference crystal structure have a higher energy. In order to compensate for this a common approach is to build models from a set of N best scoring snapshots such as introduced by Mirjalili et al. (2014). Indeed, a post-mortem analysis as seen in Figures 4.8 and 4.9 of the effect of the number of snapshots used for building models showed that large numbers are required to have an average positive refinement success on GDTHA and RMSD metrics. Relying on these large number of snapshots limits the magnitude of the theoretical possible improvements. In order to address this problem of snapshot selection a time and space dependent snapshot selection model with deep recurrent neural networks is discussed in Chapter 6.

4.4.3 Model Building Performance

In CASP11 different model building strategies were tested. Methods APR, ADR and ANR make use of averaging over the best ranked 10 percent snapshots of their respective sampling method (PR, DR and NR) to generate one final model. The other two strategies tested are M3C and MDR. Here, MODELLER is used to combine the three models APR, ADR and ANR into one model whereas in MDR the best 5 ranked snapshots from DR were used as templates for MODELLER to generate the final model. The overall success-rate to generate an improved model from one of the 5 methods is 0.833. The method with the highest individual success-rate is MDR with 0.567, followed by APR with 0.533, M3C with 0.500, ADR with 0.467 and ANR with 0.367. Interestingly, MDR which is based on the best 5 snapshots has a 0.10 higher success-rate compared to ADR which uses 10 percent of the best ranked snapshots. Furthermore, the median improvement for MDR is notably better compared to ADR with ΔGDTHA values of 0.004 and -0.003, respectively. This shows that MODELLER has better capabilities of generating improved models from a set of snapshots. However, in order to fully

quantify this, exhaustive testing would be necessary with different numbers of snapshots and snapshots from different sampling strategies in order to find the best model building strategy.

In CASP12, only averaging of snapshots for model building was considered. The methods 20CSCME, 20CSCMM, 20DCME and 20DCMM used the best 20 scoring snapshots whereas the reference method APR used the best 10 percent scoring snapshots. The overall success-rate in this round to generate an improved model in one of the 5 generated models is 0.583. The success-rate for APR dropped from 0.533 in CASP11 to 0.167 in CASP12, showing the increased difficulty of the refinement targets.

An issue with model building from averaging over an ensemble of snapshots, as done for strategies ANR, ADR, APR, 20CSCME, 20CSCMM, 20DCME and 20DCMM, is the introduced un-physical distortion of side-chains. In the presented work steepest decent energy minimization was used to resolve this problem. An alternative solution to averaging could be the selection of one centroid snapshot from the ensemble that represents the final model. This would avoid issues that could arise from the minimization process such as a non-convergence.

CHAPTER 5

Predicting the Unbound to Bound Conformational Change of Protein-Protein Complexes

5.1 Introduction

The vast majority of all proteins act as part of complexes or large assemblies where they form stable complexes with one partner, or more often have transient interactions with a large number of different partners. Resolving the three-dimensional description of these interactions at atomic detail is crucial for understanding biological function (Jones and Thornton, 1996; Nooren and Thornton, 2003). However, the number of resolved structures of protein-protein complexes in the Protein Data Bank (PDB) remains limited despite the ever-increasing number of new structures (Marsh and Teichmann, 2015). This limits the progress of our understanding of the workings of protein-protein interactions. Thus, in-silico predictions of protein-protein interactions seem to be the only viable option to complete the missing links in the structural interaction network.

Several protein-protein docking methods have been developed to predict the three-dimensional interaction of proteins and can be grouped into rigid body (Eisenstein and Katchalski-Katzir, 2004; Comeau et al., 2004; Mandell et al., 2001;

Tovchigrechko and Vakser, 2006; Schneidman-Duhovny et al., 2005; Terashi et al., 2007; Chen et al., 2003) and flexible docking (Zacharias, 2003; Dominguez et al., 2003; Fernández-Recio et al., 2003; Lyskov and Gray, 2008; Moal and Bates, 2010) methods. The former considers only translational and rotational search whereas the latter also incorporates conformational flexibility into the docking process in order to model conformational transitions from unbound to bound states. The rigid body case in docking, where the unbound structure is equal to the bound structure, is considered solved today and many highly optimized algorithms based on fast-Fourier transformation (FFT) techniques and geometrical hashing have been proposed. However, these methods fail to find high-resolution models when the proteins undergo complex conformational changes from unbound to bound.

In order to model side-chain and backbone rearrangements a high number of degrees of freedom have to be considered, requiring heuristic optimization algorithms to search the solution space efficiently. The CAPRI-experiments (Critical Assessment of PRediction of Interactions) have shown that heuristic methods are often able to find solutions with acceptable quality. Though, finding medium or high quality solutions still remains challenging. A solution to this problem are so called refinement methods which perform a local optimization of a docked solution in order to obtain higher quality models. Physics based refinement methods using molecular dynamics simulation have shown anecdotal success of improving docking solutions. However, the computational cost involved simulating long enough time scales to escape local minima has often been a limiting factor.

Here, a method is presented that exploits a so called contact map space (CMS) definition in order to perform more directed refinement compared to standard MD methods. The CMS is constructed from the observed residue-residue contacts at the interface between a receptor and a ligand from an initial docked solution. This CMS is used as a collective variable (CV) in a metadynamics simulation in order to bias the potential. From these simulations the conformational free energy landscape is reconstructed and a new scoring function CS_α is proposed that combines empirical terms from ZRANK with the reconstructed conformational free energy in order

to achieve better performance at identifying snapshots with improvements in the trajectory.

5.2 Methods

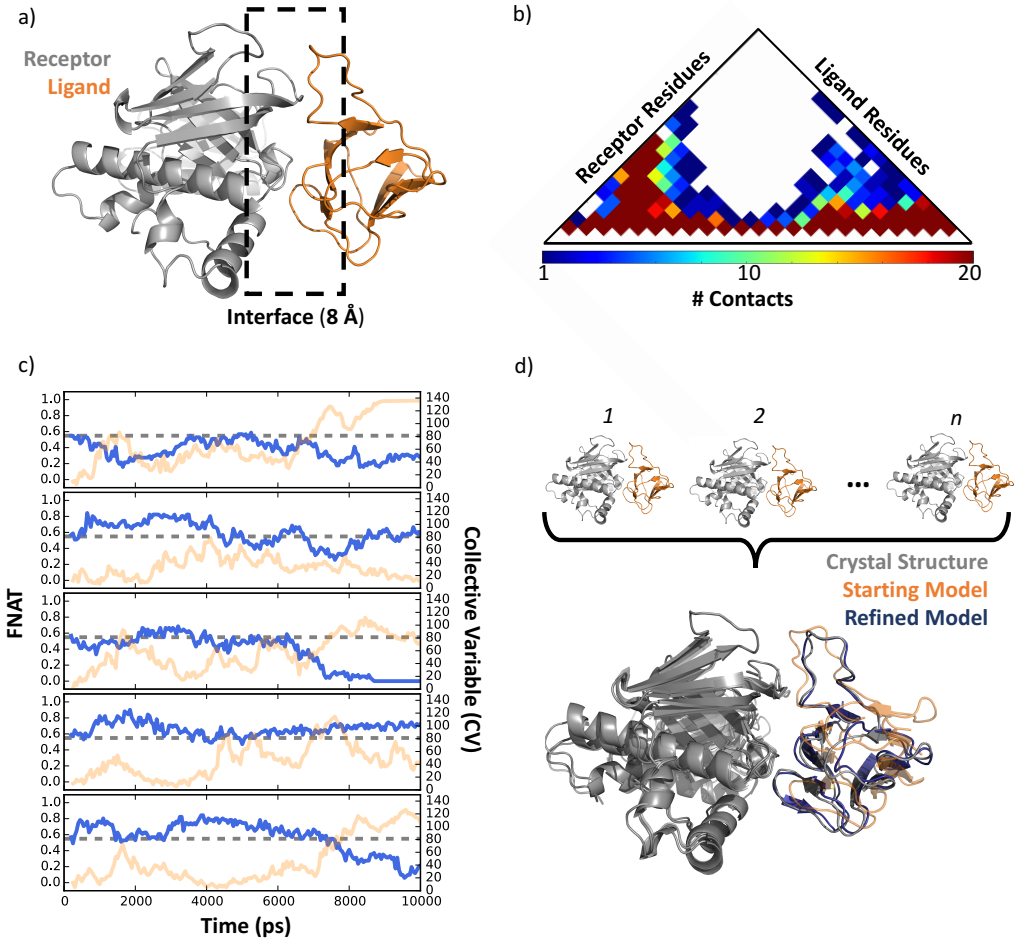

Figure 5.1: Schematic overview of the protein-protein refinement method. a) schematic representation of the interface definition which includes residues which are within 8 Å of the receptor-ligand. b) schematic representation of the contact map (CM) resulting from the residue-residue contacts at the receptor-ligand interface. The number of observed contacts comes from the ensemble of docked solutions, thus, can be greater than one. c) Example of a five times replicated sampling run of a target with metadynamics where the CV describes the CMS. The blue line represents the FNAT, the orange line the CV and the dotted gray line is the starting model FNAT. d) Selection of the best N scoring snapshots from the trajectories. The final model is an average of all snapshots by averaging the Cartesian coordinates of each atom followed by a two step energy minimization of the structure.

5.2.1 Dataset

The protein-protein refinement method was benchmarked on 23 cases using 11 targets from the score_set dataset (Lensink and Wodak, 2014) of the CAPRI scoring experiment. The dataset consists of decoys of varying quality (high, medium, acceptable and incorrect) from all participating groups. Targets containing more than one chain for receptor or ligand (T37 and T50) and targets without any acceptable, medium or high quality solutions (T36 and T38) were removed from the benchmark set (see Table 5.1 for the full list). The structure chosen to represent a quality category of a target was the centroid element based on LRMSD for all models belonging to this category. Table 5.1 gives an overview of all starting models with their initial LRMSD, IRMSD and FNAT.

5.2.2 Definition of the Contact Map Space

The contact map space (CMS) for protein-protein complexes describes the interface contacts between a receptor protein and a ligand protein. In order to qualify as a contact the distance between the Cα or the Cβ residue has to be below 8 Å. The contact map (CM) based on these contacts is named CM_{if}. The CM_{if} is described as follows:

$$CV(R) = 1/N \sum_{\gamma \in CM_{if}} (D_\gamma(R) - D_\gamma(R_{\text{ref}}))^2 \tag{5.1}$$

$$D_\gamma(R) = \frac{1-(r_\gamma/r_\gamma^0)^n}{1-(r_\gamma/r_\gamma^0)^m} \tag{5.2}$$

The sigmoid distance function $D_\gamma(R)$ quantifies the formation of a contact γ in structure R, where r_γ is the contact distance in structure R and r_γ^0 is the contact distance in reference structure R_{ref}. Here, R_{ref} describes a set of models of a target that have the same starting model quality as the selected starting model. Variables

Table 5.1: CAPRI starting model quality. Shows the FNAT, IRMSD, and LRMSD to the reference crystal structure for 23 different starting models from 11 different protein targets. The column SMQ describes the CAPRI starting model quality as assigned in the score_set dataset with the 3 classes acceptable (acc), medium (med) and high (hig).

TR	SMQ	FNAT	IRMSD (Å)	LRMSD (Å)
T29	acc	0.45	3.41	6.98
T29	hig	0.82	1.82	3.83
T29	med	0.53	2.75	5.21
T30	acc	0.20	6.12	13.13
T32	acc	0.36	2.77	8.08
T32	med	0.49	1.96	6.57
T35	acc	0.15	5.09	13.30
T39	acc	0.55	2.31	7.51
T39	med	0.78	1.32	3.65
T40	acc	0.63	2.58	6.84
T40	hig	0.80	1.03	4.32
T40	med	0.80	2.16	4.27
T41	acc	0.49	2.63	6.97
T41	hig	0.78	0.80	2.48
T41	med	0.65	1.38	3.40
T46	acc	0.49	3.75	10.57
T47	acc	0.54	2.56	5.70
T47	hig	0.85	0.99	1.59
T47	med	0.79	1.32	2.84
T53	acc	0.19	5.67	13.09
T53	med	0.48	5.70	9.62
T54	acc	0.41	3.94	7.53
T54	med	0.50	2.70	4.76

n and m are constant and set to $n = 6$ and $m = 10$.

5.2.3 Simulation Setup

All starting models were checked for missing residues and atoms, and when necessary were completed with the program Loopy (Xiang et al., 2002) and SCRWL (Krivov et al., 2009). The system was solvated in a cubic simulation box with a buffer of 12 Å using the explicit solvent model SPC (Jorgensen et al., 1983) and the charge was neutralized with Na^+ and Cl^- ions with a concentration of 0.15 mol/liter. The energy minimization was performed with GROMACS 4.6 and

consisted of the following three steps: i) steepest decent energy minimization with 50000 steps and a step-size of 0.01; ii) conjugate gradient based minimization with 500000 steps and one steepest decent step every 1000 steps; iii) a second round of steepest decent minimization for 50000 steps. The equilibration of the system, using GROMACS 4.6, followed a two step protocol where all heavy atoms were subject to position restraints with a force of 1000 kJ $mol^{-1}nm^{-1}$. The first phase consisted of an 100 ps long NVT equilibration where an increase of the temperature with V-rescale (Bussi et al., 2007) from 0 K to 300 K has been performed. In the second step, a NPT equilibration was performed where the pressure of the system is increased to 1 bar with Parrinello Rahman pressure coupling (Berendsen et al., 1984) for a simulation time of 300 ps. The production run with metadynamics in CMS as defined in Equation (5.1) was performed with PLUMED2 and GROMACS 4.6. The Gaussian addition is deposited every 2 ps with $\sigma = 0.5$ and a bias factor of 10 and an initial height of 5 kJ mol^{-1}. The sampling was performed for 10 ns and snapshots were saved every 2 ps.

5.2.4 Definition of the Scoring Function CS_α

The new scoring function CS_α combines the *FES* reconstructed from metadynamics simulations with the ZRANK scoring function as follows:

$$CS_\alpha = \alpha ZRANK_N + (1-\alpha) FES_N \tag{5.3}$$

The two functions FES_N and $ZRANK_N$ are the 0 to 1 normalized energies. The parameter α is a weighting factor that ranges from 0 to 1. For example, an α-value of 1 means that only $ZRANK_N$ is considered for the scoring and a value of 0 means that only the FES_N is considered.

The correct rank for a set of snapshots s for each target tr is given by the ascending order of their LRMSD to the reference crystal structure. This sorted list of snapshots is defined as $sort_{lrmsd}(s)$. Furthermore the maximum rank is capped such that

$$rank(s) = \begin{cases} i, & \text{if } i \leq max \\ max, & \text{otherwise} \end{cases} \tag{5.4}$$

where max is the threshold that is used when $i > max$. Applying these two functions to s gives the reference ranking $r = rank(sort_{lrmsd}(s))$. The rank assignment based on function CS_α is the descending order of their scores and the ranks produced by this function is denoted as $c = rank(sort_{cs}(s))$. Following this notation, the rank for snapshot i is retrieved by r_i and c_i, respectively. The ranking error ε produced by CS_α can now be quantified with

$$\varepsilon = \sum_{tr} \sum_{i=0}^{c:N} r_i, \tag{5.5}$$

where N is the number of snapshots that are used for ranking. And is normalized to

$$\varepsilon_N = \frac{\varepsilon - rank_{min}}{rank_{max} - rank_{min}}, \tag{5.6}$$

where $rank_{min} = |TR|((N(N+1))/2)$ and $rank_{max} = |TR|(max+1)N$.

5.2.5 Model Building

Model building is based on the best N ranked snapshots from FES_N, $ZRANK$ and CS_α ($\alpha = 0.49$) where the three final models are named AFES, AZRANK and ACS. The final model is computed by averaging each atom's coordinates from the N selected snapshots from a target's trajectory. Snapshots with a $\Delta t = 50$ are considered for model building. An energy minimization of the averaged model with steepest decent and 50000 steps was performed in order to resolve non-physical conformations.

5.2.6 Model Assessment Measures

The model quality is assessed by the three metrics LRMSD, IRMSD and FNAT. A detailed definition of these model assessment measures is given in Section 2.6.2.

5.3 Results

Table 5.2: Complex model quality after refinement. The table shows the results from 11 different target complexes (TR) with different CAPRI starting model qualities (SMQ) acceptable (acc), medium (med) and high (hig). Metrics shown are ΔFNAT, ΔLRMSD (Å) and ΔIRMSD (Å) from model building with the best 14 snapshots as selected by $ZRANK_N$ and for the best snapshots with $\Delta t = 50$ as generated during the sampling. Values in bold indicate an improvement over the initial model quality.

		Build Model with N=14			**Best Snapshot**		
TR	**SMQ**	**ΔFNAT**	**ΔLRMSD**	**ΔIRMSD**	**ΔFNAT**	**ΔLRMSD**	**ΔIRMSD**
T29	acc	**0.08**	**-1.66**	**-0.90**	**0.24**	**-4.25**	**-1.59**
T29	med	-0.04	1.61	0.70	**0.16**	**-2.35**	**-0.38**
T29	hig	-0.04	**-0.10**	**-0.37**	0.00	**-1.15**	**-0.22**
T30	acc	**0.09**	1.75	**-0.25**	**0.25**	**-3.55**	**-0.75**
T32	acc	**0.22**	**-4.75**	**-1.13**	**0.24**	**-4.93**	**-1.53**
T32	med	-0.01	**-3.25**	**-0.43**	**0.16**	**-3.48**	**-0.84**
T35	acc	-0.06	0.10	0.35	**0.02**	**-5.04**	**-0.87**
T39	acc	**0.24**	**-5.92**	**-1.28**	**0.35**	**-6.03**	**-1.76**
T39	med	**0.16**	**-1.50**	**-0.10**	**0.18**	**-2.35**	**-0.50**
T40	acc	0.00	**-1.29**	**-0.60**	**0.11**	**-4.62**	**-0.70**
T40	med	-0.05	**-0.27**	**-0.24**	**0.02**	**-2.20**	**-0.62**
T40	hig	-0.08	4.58	1.30	**0.13**	**-2.24**	**-0.15**
T41	acc	-0.13	1.12	1.75	**0.04**	**-3.57**	**-0.42**
T41	med	-0.25	1.50	1.64	**0.07**	**-1.13**	0.12
T41	hig	-0.31	1.96	1.54	**0.03**	**-0.92**	0.09
T46	acc	-0.03	0.03	**-0.30**	**0.01**	**-3.00**	**-0.08**
T47	acc	**0.06**	0.91	**-0.21**	**0.17**	**-2.76**	**-1.12**
T47	med	-0.13	2.87	0.77	**0.02**	**-1.34**	**-0.22**
T47	hig	-0.10	1.93	0.46	**0.06**	**-0.07**	**-0.01**
T53	acc	**0.17**	**-0.69**	0.97	**0.29**	**-2.45**	**-0.84**
T53	med	**0.04**	0.92	**-1.45**	**0.33**	**-3.20**	**-1.01**
T54	acc	**0.09**	**-1.78**	**-1.21**	**0.19**	**-3.22**	**-1.64**
T54	med	0.00	**-0.13**	**-0.27**	**0.09**	**-1.39**	**-0.37**

5.3.1 Overall Refinement Success

The results shown in this section give a general overview of the refinement success with snapshot selection based on selecting the best 14 ranked snapshots with $ZRANK_N$ for model building and a comparison to the best generated snapshot (see

Table 5.2). The selected number of snapshots is taken from the analysis in section 5.3.4 where N=14 produced the best average refinement success based on ΔFNAT for models with acceptable starting model quality.

Overall, the refinement was mostly successful in the acceptable category, here the FNAT, LRMSD and IRMSD could be improved for 7, 6 and 8 out of 11 targets, respectively. For models starting from medium quality the metric for FNAT, LRMSD and IRMSD could be improved 2, 4, 5 out of 8 targets, respectively. For the 4 high quality examples the FNAT, LRMSD and IRMSD was improved 0, 1, 2 times, respectively. These results suggest that the method has a modest success at improving the interface region. The most successful refinement was possible for target T39. Starting from an acceptable quality model where FNAT, LRMSD and IRMSD improved by 0.24, -5.92 Å, -1.28 Å from the initial values 0.55, 7.51 Å, 2.31 Å, respectively.

The theoretical best refinement success, if the best snapshot would have been selected as the final model, yields good results for all three starting model quality classes. The FNAT, LRMSD and IRMSD could be improved for all acceptable quality models. The sampling for medium starting quality failed only for target T41 where the IRMSD decreased slightly by 0.12 Å whereas all other metrics in all other targets could be improved. Similarly, for starting models with high quality, here a decrease in quality could be only observed for the IRMSD of T41 with 0.09 Å. The target with the largest improvement for FNAT, LRMSD and IRMSD is T39 starting from acceptable quality with an improvement of 0.35, -6.03 Å and -1.76 Å, respectively.

5.3.2 Refinement Success as a Function of Time

The analysis of the sampling power for the 5 times replicated metadynamics runs in CMS for 10 ns shows that large FNAT and LRMSD improvements are mostly sampled within the first 4 ns, where snapshots with improvements of $\Delta\text{FNAT} > 0.25$ and $\Delta\text{LRMSD} < -4.5$ have the highest density (see Figure 5.3, panel left and centre). Snapshots with smaller FNAT improvements (0.01 to 0.1) have a uniform

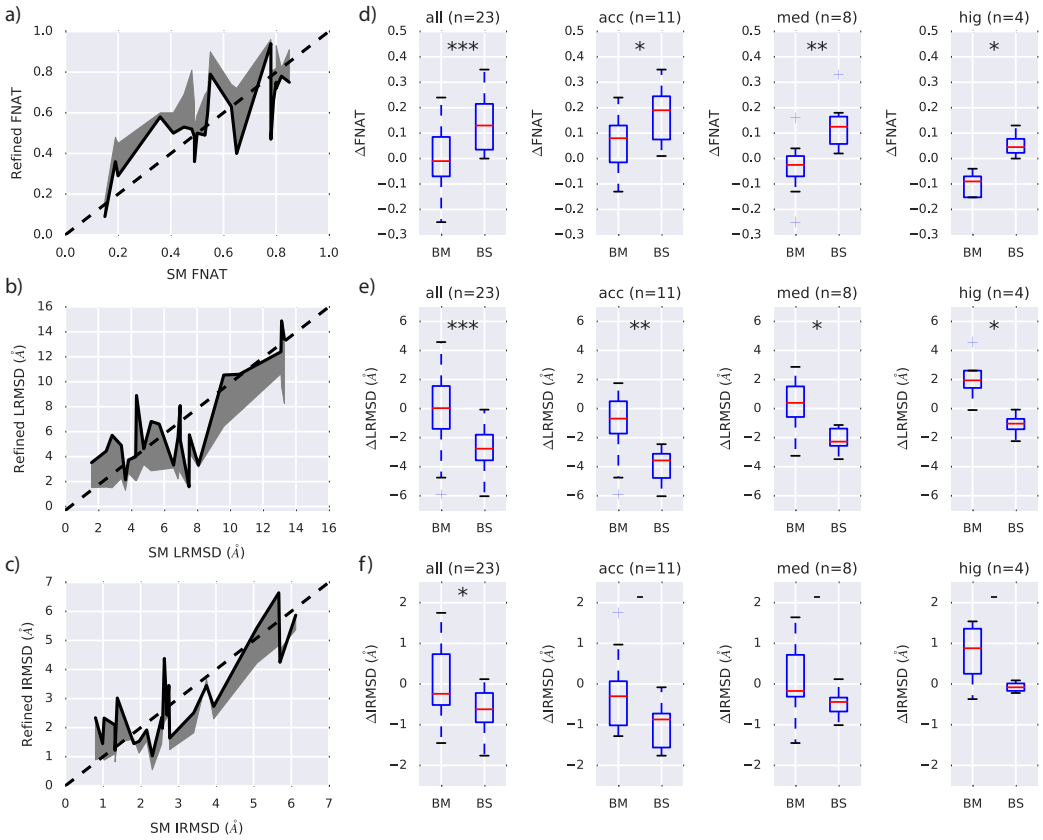

Figure 5.2: Complex refinement result overview. a) Starting model (SM) FNAT versus refined FNAT. b) Starting model (SM) LRMSD versus refined LRMSD. c) Starting model (SM) IRMSD versus refined IRMSD. For a), b) and c) the black line indicates the change based on build model and the gray area based on best snapshot. d), e) & f) Split down of refinement results with respect to starting model quality all, acceptable (acc), medium (med) and high (high). The number n in brackets indicates the number of refined models in that category. The symbols ***, **, * and - indicate significance level between build model (BM) and best snapshot (BS) at p-value < 0.001, p-value < 0.01, p-value < 0.05 and p-value ≥ 0.05, respectively.

distribution with equal density across sampling time, whereas for LRMSD the density continuously lowers with increasing sampling time for all shown thresholds.

Improvements for large IRMSD deviations (> -1.4 Å) follow a bimodal distribution where the highest density of these snapshots are observed around time-points 2 ns and 8 ns as seen in Figure 5.3, right panel. The smaller IRMSD improvements follow a uniform density distribution across the whole sampling time (thresholds -0.6 Å to -1.0 Å). This could indicate that transitions to larger IRMSD improvements would possibly require longer simulation time.

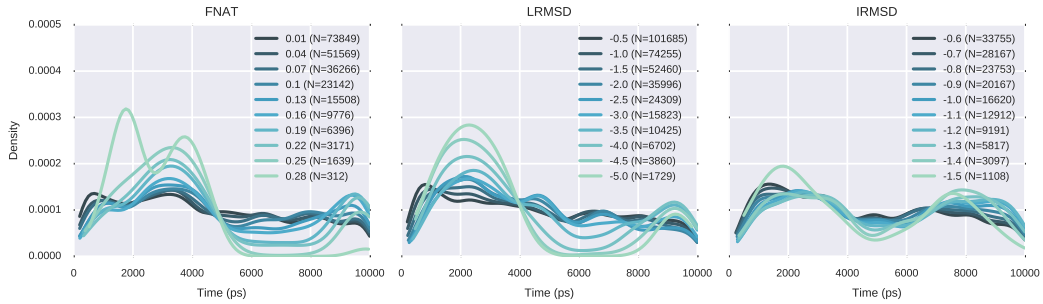

Figure 5.3: Complex refinement improvements as a function of time. Shown is the improvement over time for FNAT, LRMSD (Å), IRMSD (Å). The different coloured lines show the used threshold. The number N in brackets indicates the number of snapshots $\geq$ the threshold.

5.3.3 Snapshot Selection with CS_α

The results in Table 5.2 show a large difference between the final model generated by the best 14 ranked snapshots as selected by $ZRANK_N$ and the theoretical best improvement if the best snapshot would have been selected. In order to improve the snapshot selection the scoring function CS_α was constructed where $ZRANK_N$ and FES_N are combined by a weighting factor α. In a first step the effect of different α-values on the ranking error ε with respect to different number of selected snapshots N is explored. The heatmap in Figure 5.4a shows that if only the best snapshot is selected ($N = 1$) a value of $\alpha = 1.0$ where only $ZRANK_N$ is contributing to the ranking will result in the lowest ranking error. However, with increasing N the contribution of FES_N becomes more important as indicated by the gray line in Figure 5.4a and converges to a stable value of $\alpha = 0.49$. Furthermore, the heatmap also shows that high contributions of FES_N with $\alpha < 0.2$ leads to high ranking errors.

An example of scoring the different snapshots versus FNAT is shown in Figure 5.4b for target T39 starting from an acceptable quality model. The left panel shows that FES_N has a broad energy funnel where snapshots with a wide range of FNAT values (0.5-0.9) have similar energies. Thus, making a selection of the best snapshots impossible. The funnels of $ZRANK_N$ and $CS_{0.49}$ show a better correlation with FNAT and are similar for the shown target, where snapshots with higher FNAT values yield higher scores. In Figure 5.4c the effect of the number

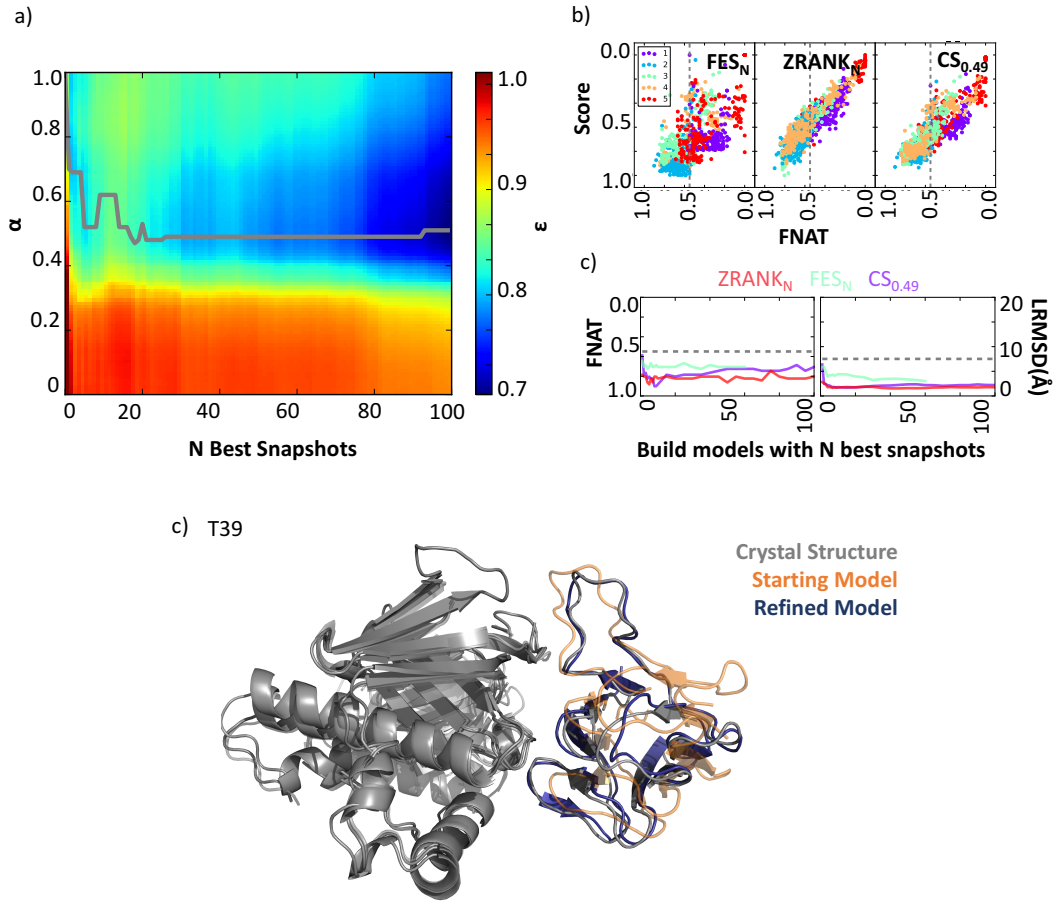

Figure 5.4: Parameter optimization of CS_α. a) Heatmap showing ranking error ε for different α and number of N selected snapshots, b) Comparison of scoring functions FES_N, $ZRANK_N$ and $CS_{0.49}$ with respect to FNAT. The scatterplot shows snapshots (Δt = 50 ps) for 5 replicated sampling runs of target T39 starting from an acceptable solution. c) shows the change of FNAT (left) and LRMSD (right) after model building by selection the N best snapshots as ranked according to $ZRANK_N$, FES_N and $CS_{0.49}$. c) Shows a 3D rendering of target T39 before (orange) and after (blue) refinement, all models are superimposed to the crystal structure (gray).

of snapshots selected with $ZRANK_N$ FES_N and $CS_{0.49}$ with respect to FNAT (left panel) and LRMSD (right panel) after model building for target T39 is shown. All three functions are able to improve the model quality for N ranging from 1 to 100. However, function FES_N achieves smaller improvements compared to $ZRANK_N$ and $CS_{0.49}$. The two later functions have equal performance with respect to LRMSD improvements, whereas the snapshot selection with $ZRANK_N$ produces better FNAT models for snapshot ranges 40-100.

5.3.4 Optimizing for the Number of Snapshots

A more detailed analysis of the effect of N on model building is shown in Figure 5.5. Function $ZRANK_N$ is able to produce on average improved models for starting models with acceptable quality for a wide range of N (Figure 5.5, left columns and blue labels). Where the best results for average ΔFNAT, ΔLRMSD and ΔIRMSD was achieved at $N = 14$ with 0.07, $N = 16$ with -1.16 Å and $N = 61$ with -0.41 Å, respectively. For medium quality starting models (left columns and green labels) the average quality decreased with a value of ΔFNAT$= -0.01$ ($N = 18$) and increased slightly with ΔLRMSD=-0.25 Å ($N = 189$) and ΔIRMSD=-0.18 Å ($N = 193$). As expected, high quality starting models could not be improved on average. All three metrics decreased in quality after refinement (left columns and red labels). The best values were obtained for FNAT, IRMSD and LRMSD at $N = 7$ with -0.11, $N = 7$ with 1.74 Å and $N = 109$ with 0.59 Å, respectively.

Centre columns of Figure 5.5 show the results for FES_N. The average ΔFNAT for acceptable, medium and high starting models could not be improved and yielded the best results at N=196 with -0.10, N=73 with -0.10 and N=70 with -0.34, respectively (Figure 5.5a). For ΔLRMSD, only acceptable models could be improved where the best N=193 had a ΔRMSD of -0.39 Å. Medium and high quality models decreased on average in quality by 0.96 Å (N=129) and 3.82 Å (N=70), respectively (Figure 5.5b). Similar results are observed for ΔIRMSD, as shown in 5.5c, where only acceptable starting models could be improved with by -0.486 Å (N=189). The ΔIRMSD decreased for medium and high by 0.01 Å (N=191) and 1.25 Å (N=70).

The newly introduced scoring function $CS_{0.49}$, shown in the right columns of Figure 5.5, did not improve model building performance compared to $ZRANK_N$, and is only on par for acceptable starting models. Here, the best average ΔFNAT, ΔLRMSD and ΔIRMSD improvement was yield at N=20 with 0.06, N=38 with -1.11 Å and N=93 with -0.38 Å, respectively. For medium quality starting models the ΔFNAT decreased at best N=162 with -0.05 and improved for ΔLRMSD and ΔIRMSD for best N=135 with -0.06 Å and N=194 with -0.14 Å, respectively.

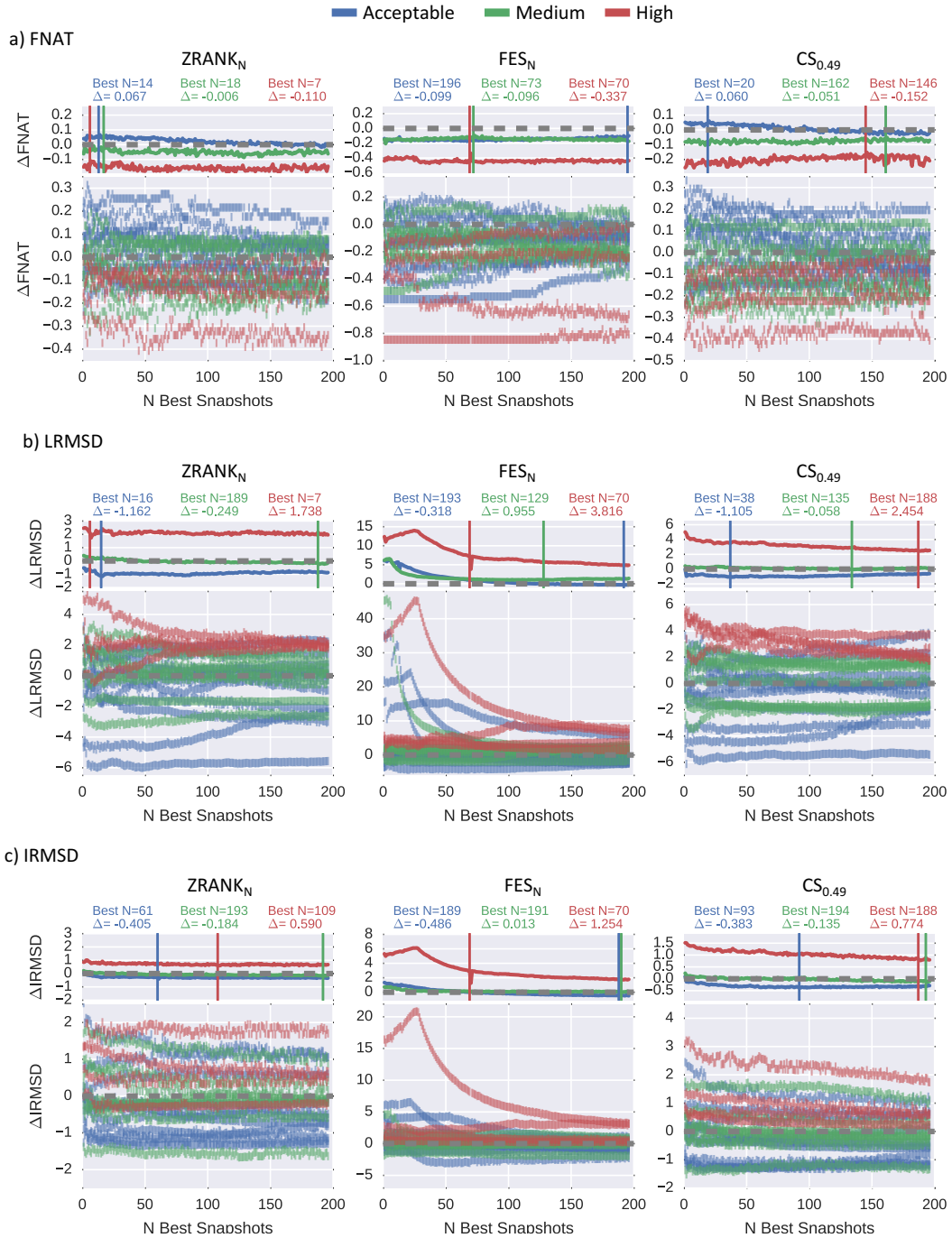

Figure 5.5: Optimization of the number of snapshots for model building. a) ΔFNAT b) ΔLRMSD (*Å*), c) ΔIRMSD (*Å*). The three colors blue, green and red represent the target's starting model quality acceptable, medium and high, respectively. The grey dotted line indicates the Δ-value at 0.0. The three lines in the upper row for each subplot a, b and c show the average value grouped by the three different starting model qualities. The horizontal lines indicate the best N based on best average performance for the metric, *N* and Δ value are shown on top. The scatter plot in the bottom row of each subplot a, b and c shows the actual metric's value for each target.

Finally, for high quality models all three metrics ΔFNAT, ΔLRMSD and ΔIRMSD decreased after refinement for all N tested snapshots. The smallest decrease in quality was yield at N=146 with -0.15, N=188 with 2.45 Å and N=188 with 0.77 Å, respectively.

5.4 Discussion

5.4.1 Increased Sampling Power with more Replicated and Shorter Runs

The results show that sampling in CMS with metadynamics yields improved quality snapshots for all targets and starting model quality categories. Sampled improvements for FNAT ranged from 0.01 to 0.35, for LRMSD from -0.07 Å to -6.03 Å and for IRMSD from -0.01 Å to -1.76 Å. Interestingly, the largest improvements for FNAT and IRMSD were mainly sampled in the first 4 ns of the refinement runs, suggesting that shorter and more replicated runs could lead to enhanced sampling power for those two metrics. An explanation for this finding is the observation that during the sampling runs disassociations between the receptor and ligand can occur and hence resulting in solutions with high LRMSD and 0 FNAT. A solution to this problem could be the introduction of a so called upper wall defined by an LRMSD to the starting configuration that introduces an energy barrier when the LRMSD becomes greater than a pre-defined value and thus preventing disassociation events from occurring during the sampling process. This was successfully tested for target T54 where the receptor-ligand complex stayed in contact for the whole simulation time. However, a complete evaluation for the full benchmark set was not performed.

5.4.2 Sampling of Unbound to Bound Conformational Transitions

The main focus of the presented method was to improve directed sampling at the interface level. The metric IRMSD quantifies the conformational difference

at the interface between the predicted model and the reference crystal structure state. Table 5.1 shows the IRMSD of the used starting models, where values range from 0.80 Å to 6.12 Å. The results in column "Best Snapshot" of Table 5.2 and Figure 5.3 show that substantial improvements could be sampled for a number of targets. For example, T29 improved with an ΔIRMSD of -1.59 Å from an initial value of 3.41 Å. However, sampling full transitions, where IRMSD values below 1 Å are obtained, remains challenging. This was already observed in a study by Kuroda and Gray (2016), where all tested sampling methods failed to fully sample the full transition from an unbound to bound conformational state. Nevertheless, the presented approach by sampling in CMS remains promising. A future optimisation to the sampling process could be obtained by testing different values for the metadynamics parameters bias factor and initial height. The results shown here are based on a bias factor of 10 and an initial height of 5 kJ mol^{-1}. A higher value for the bias factor could increase the sampling around the defined CMS leading to more effective sampling at the interface. Whereas higher values for the initial height results in smaller energy barriers between local minima in the FES. The effect of this would be faster exploration of different states along the defined CMS. However, too high values of initial bias and height could lead to sampling of non-relevant high energy states. Systematic parameter testing on the benchmark set of these two parameters has to be performed in order to find the optimum values for efficient and successful sampling on a wide range of different protein complexes.

5.4.3 Model Building Procedure is Successful at Generating Improved Docked Models

The results for model building based on snapshot selection with $ZRANK_N$ and $CS_{0.49}$ have shown some degree of success for starting models with acceptable quality, where the FNAT, LRMSD and IRMSD could be improved 7, 6 and 8 times out of 11 targets. If a success is defined as improving at least one metric a success rate of 81.82 percent could be achieved where 9 out of 11 targets are improved. It has been shown that FES_N alone produces a higher ranking error of snapshots

compared to $ZRANK_N$ and thus, is not recommended as a viable alternative for snapshot selection. The function $CS_{0.49}$ is comparable in performance to $ZRANK_N$ for improving acceptable quality models but falls behind the performance when medium quality or high quality models are refined. Thus, snapshot selection solely based on $ZRANK_N$ is recommended. Furthermore, the results also show that model building is stable for a large number of different ranges of snapshots (see Figure 5.5). This means that positive refinement success can be expected for a large number of N.

5.4.4 Future Directions

Currently, the definition of the CMS is based on observed residue-residue contacts between a receptor and a ligand from one or a set of initially docked poses in the score_set dataset. Alternativley, the CMS could be defined from inferred residue-residue contacts from sequence co-evolution analysis as shown by Ovchinnikov et al. (2014) and Hopf et al. (2014). They outline a methodology how correlated mutations from large multiple sequence alignments between the receptor and ligand protein sequence can identify residue-residue pairs close in space. A CMS based on this data is not restricted to the solutions based on initial docking results but allows for a wider range of contacts. Refinement based on this CMS definition would not only allow biased sampling along these putative contacts, but also enables the reconstruction of a residue-residue energy landscape that describes different contacts. Such an approach, combining co-evolution information and a physical description of these energy landscapes, could potentially lead to 3D protein-protein interaction models with high accuracy.

The snapshot selection tested in this chapter is based on computed energies from $ZRANK_N$, FES_N and $CS_{0.49}$ alone. However, the sampling method presented in this work produces trajectories where the exploration of a systems configuration R at time-point t denoted as R_t depends on the previous configuration of the system at time point R_{t-1}. Thus, a function which evaluates the quality of the snapshots based on temporal dependencies of energies in time could lead to a better snapshot

selection. Recent advances in deep learning have shown that such temporal (also known as sequence dependencies) can be modelled with recurrent neural networks (Hochreiter and Schmidhuber, 1997). A detailed description of such a model is presented in Chapter 6 for protein monomer refinement. However, an adaptation for complex refinement trajectories should be easily possible.

CHAPTER 6

Learning to Predict Improved Conformations of Proteins with Deep Recurrent Neural Networks

6.1 Introduction

In Chapter 4 a refinement method to optimize the model quality of predicted protein structures was presented. This method exploited MD and metadynamics simulations to sample the conformational space in order to explore configurations that have improved model quality with respect to a reference crystal structure. It was shown that such improved configurations could be sampled. However, reliable identification of improved quality configurations from these is challenging. The previous results of Chapter 4 showed that the theoretical best improvement in ΔGDTHA, by picking the best snapshot, is markedly better than the ΔGDTHA improvement identified with the knowledge-based scoring function DFIRE. The focus of research in MD based refinement has so far neglected the temporal dimension in snapshot selection, where the configuration of a protein system at time-point t is dependent on its previous state at t-1 and its negative gradient vector of a potential energy function.

Thus, in this chapter a spatial-temporal model for snapshot classification of

MD trajectory data is formalized that makes explicit use of the time-dependent nature of MD based trajectory data. In particular, the interest lies in whether it is possible to identify when, or if improved quality conformations of a protein are reached. From such a trajectory of snapshots, predictions about a protein's conformational states are based on energies and distance metrics in time. To this end a deep recurrent neural network (RNN) with gated recurrent units (GRU) is trained to classify each snapshots into one of the three classes: improved quality, no-change in quality, and decreased quality snapshot, where the change of quality is defined as an increase or decrease in GDTTS from the starting configuration as measured with respect to the reference crystal structure.

The results show that it is possible to train a RNN model that identifies improved and decreased quality poses. Furthermore, the here proposed model outperforms classic machine learning models such as random forest (RF), logistic regression (LR) and k-nearest neighbours (KNN). This further proofs that the temporal patterns learned by the RNN are important and contribute to a higher precision of identifying improved quality snapshots.

6.2 Methods

6.2.1 Model Definition and Training

The model learns via a supervised learning-task to assign the class y from a set of given input features, x, for each time-point, $[\tau]_i$, of a trajectory ν. Here, the three possible classes are improved, no-change and decreased. The ground truth assignment, denoted as y', is then formalized such that

$$y' = \begin{cases} \boldsymbol{i} & \text{if } \Delta\text{GDTTS} \geq 0.01 \\ \boldsymbol{n} & \text{if } \Delta\text{GDTTS} < 0.01 \text{ and } \Delta\text{GDTTS} \geq -0.01 \\ \boldsymbol{d} & \text{if } \Delta\text{GDTTS} < -0.01, \end{cases} \tag{6.1}$$

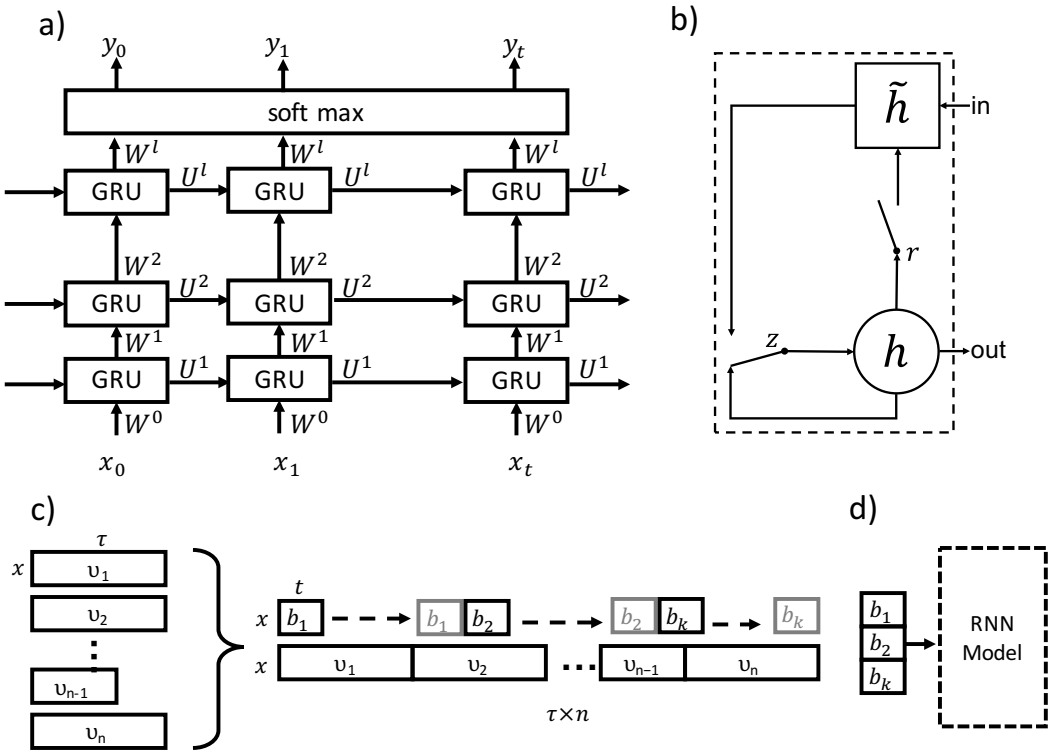

Figure 6.1: RNN model description. a) Schematic overview of the RNN with GRU cells, a detailed explanation is given in Section 6.2.1. b) GRU cell, visualisation of Equations 6.3, 6.4, 6.6 and 6.5. c & d) Visualisation of the trajectory data v and the process of mini-batch creation and propagation to the RNN.

where $\boldsymbol{i}$, $\boldsymbol{n}$ and $\boldsymbol{d}$ represent one-hot encoded probability vectors, where one element of the vector is 1 and the rest 0, $[1,0,0]$, $[0,1,0]$ and $[0,0,1]$, respectively. The variable ΔGDTTS is the difference between GDTTS values from starting model versus crystal structure and snapshot model versus crystal structure, where a negative Δ reflects an decrease in model quality and a positive Δ an increase in model quality. A detailed definition of GDTTS is provided in Section 2.6.1.

The model to learn the given task is based on a RNN with GRU that adaptively learns long and short term dependencies of inputs to assign the class y (Cho et al., 2014). The layout of the RNN is illustrated in Figure 6.1a, where starting from the input sequence $x_0,\ldots,x_t$ the predictions $y_0,\ldots,y_t$ are produced via stacking hidden layers of GRU cells and by a layer with a softmax activation function Ω that

normalizes the output h_t^l (at t of the last layer l) to a probability vector y such that

$$[y_t]_j = \Omega_j(h_t^l) = \frac{\exp\left(W_j^l h_t^l\right)}{\sum_{j'}^{C} \exp\left(W_{j'}^l h_t^l\right)}, \tag{6.2}$$

for all $j = 1,...,C$ classes, where W_j^l are the rows of the weight matrix of the last layer W^l. The activation and its output, h_t^l in layer l at time t, of a GRU cell is computed as

$$h_t^l = z \odot h_{t-1}^l + (1 - z_t) \odot \tilde{h}_t^l. \tag{6.3}$$

This represents a linear interpolation of the activation at time-point $t-1$ denoted as h_{t-1}^l and its candidate activation $\tilde{h}_t^l$. The update gate z controls how much the cell updates its state, such that

$$z = \sigma(W_z^l h_t^{l-1} + U_z^l h_{t-1}^l), \tag{6.4}$$

where the activation function σ is sigmoid. The candidate state $\tilde{h}_t^l$ is computed such that

$$\tilde{h}_t^l = \phi(W^l h_t^{l-1} + U^l (r \odot h_{t-1}^l)). \tag{6.5}$$

Here, ϕ, r and $\odot$ denote a hyperbolic tangent activation function, a reset gate and an element wise multiplication, respectively. Reset gate r is computed with the same formulation as z but different weight matrices, that is

$$r = \sigma(W_r^l h_t^{l-1} + U_r^l h_{t-1}^l). \tag{6.6}$$

An illustration of these equations is shown in Figure 6.1b.

During training the weight matrices W^l, U^l, W_r^l, U_r^l, W_z^l and U_z^l are learned for

each layer l. The weight matrices are shared through time t. The loss function L

$$L_{y'}(y) = -\sum_j \omega_j(y'_j \log(y_j)), \tag{6.7}$$

is minimized during training and represents the weighted cross entropy. The vector $\boldsymbol{\omega}$ encodes the weights for classes $j = 1,...,C$. The objective of this classifier is to achieve a high precision for the improved class, i.e. reducing the false positive rate and a high recall for the decreased class, i.e. reducing false negative rate. This is achieved by setting $\boldsymbol{\omega} = [0.05, 1, 10]$.

The RNN model is trained with the Adam optimizer (Kingma and Ba, 2014) on input features x from the set of trajectories ν selected for training. In order to achieve one sequential input for the training, all n trajectories are concatenated to size $\tau \times n$. The training is performed on k mini-batches b of the data where in each training iteration the batch b_k continuous without overlap to the previous batch iteration till the epoch is finished. This process is visualized in Figure 6.1c and 6.1d.

6.2.2 Data Set

The trajectory data originates from our laboratories refinement method in CASP11 and CASP12 for which the reference crystal structure is available in the PDB. The detailed description of the sampling process is described in Section 4.2.1 and 4.2.2. In total, the trajectory data consists of 3419 ns cumulated simulation time and 1709704 snapshots with $\Delta t = 2$ ps from 30 CASP11 and 12 CASP12 targets. A detailed overview of the targets is provided in in Section 2.1.1.

6.2.3 Computation of Molecular Descriptors and Feature Construction

In total 19 features were used. 17 of these features originate from ten different molecular descriptors and two features are the distance metrics GDTTS and RMSD and measure for each snapshots the difference to the starting model. Additinally, a detailed explanation of the different descriptors and distance metrics is provided in

Section 2.2.1 and 2.6.1. Furthermore, all features are normalized per target to zero mean and unit standard deviation.

6.2.4 Cross-Validation

The CV set is made up of 7 folds, where for each fold the training set consists of trajectories of 36 targets and the validation set of 6 targets. The assignment of a proteins trajectories to a fold is random. However, the relative distribution of classes of snapshots between training and validation set is enforced to be similar with a maximum difference of 6 per-cent as shown in Table 6.1 columns I, N and D. A detailed overview of each targets fold assignment is given in the supplemental material Table D.2.

Table 6.1: CASP CV summary. Summary of each fold of the 7-fold CV of the trajectory data. Shown are the number of targets (# TR), number of trajectories (# Trj), number of snapshots (# Snap.), percentage of snapshots with improved quality (I (%)), percentage of snapshots with no change in quality (N (%)), percentage of snapshots with decreased quality (D (%)).

			Training Set			
Fold	**# TR**	**# Trj.**	**# Snap.**	**I (%)**	**N (%)**	**D (%)**
0	36	780	1463580	8.38	15.00	76.62
1	36	772	1451572	8.39	14.07	77.54
2	36	772	1451572	8.54	14.00	77.46
3	36	772	1451572	7.47	14.38	78.14
4	36	780	1462780	8.09	14.46	77.45
5	36	782	1499782	8.22	14.90	76.88
6	36	766	1477366	8.35	14.87	76.78
			Validation Set			
Fold	**# TR**	**# Trj.**	**# Snap.**	**I (%)**	**N (%)**	**D (%)**
0	6	124	246124	7.16	11.74	81.10
1	6	132	258132	7.17	17.14	75.70
2	6	132	258132	6.34	17.49	76.17
3	6	132	258132	12.34	15.35	72.31
4	6	124	246924	8.90	14.93	76.17
5	6	122	209922	8.12	11.87	80.01
6	6	138	232338	7.31	12.36	80.33

6.2.5 Model Hyper-Parameter

In this work the RNN hyper-parameters sequence length, batch-size, internal size, number of layers, learning rate and dropout were systematically explored. Details of these are provided in Table 6.2. Exploration of the hyper-parameter space was performed by varying one parameter at the time from the default parameter and training and testing was performed on fold number 4 of the CV set for 300 epochs.

Table 6.2: RNN parameter. Shown are the parameter, a short description and the range of values that are adjusted during hyper-parameter testing. The values highlighted in bold represent the default parameter value.

Parameter Name	**Description**	**Range**
Sequence Length	The unrolement size of the RNN, e.g. a length of 30 makes joint predictions of 60 ps time-frames.	10, 20, **30**, 40, 50
Batchsize	Number of training examples used in each iteration in the training loop.	10, 30, **50**, 70, 100
Internal Size	The size of the internal state of the hidden units in the RNN.	256, 512 ,**1024**, 2048, 3072
Number of Layers	The depth of the RNN.	1, 2, **3**, 5, 10
Learning Rate	defines the multiplier of the derivative of the loss function during the gradient decent optimization of the weights in the network.	10^{-2}, 10^{-3}, $\mathbf{10^{-4}}$, 10^{-5}, 10^{-6}
Dropout (p-keep)	Inverse probability of setting the output of randomly selected neurons to zero.	0.7, 0.8 , **0.9**, 0.95, 1.0

6.2.6 Baseline Model

The RNN model was compared to the following baseline models:

Random Forest (Breiman, 2001): The training of the classifier uses 500 trees where samples are bootstrapped and the gini impurity criterion is used to judge the quality of the split when building the trees. No restriction for the

maximum depth of the tree is imposed , however, for each internal node in a tree the minimum sample size must be greater than 30. The number of features for each tree is $\sqrt{n}$ where n=19, i.e. the total number of features.

K Nearest Neighbour (Cunningham and Delany, 2007): Number of neighbours and the leaf-size was set to 5 and 30, respectively. A uniform distribution where all points are weighted equally in each neighbourhood was chosen. The algorithm to search for the nearest neighbours was set to 'auto' where the best algorithm from ball-tree (Liu et al., 2006), kd-tree (Bentley, 1975) and a brute force approach was selected for fitting the model with the distance metric minkowski with $p = 2$ which is equivalent to the Euclidean distance.

Logistic Regression (Bishop, 2006): Fitting of the model is performed with L2 regularization with a strength of 1.0 and with a tolerance of $1e-4$ as the stopping criteria. In order to make the multi-class predictions with LR, the training task is translated into a binary classification problem where for each label a fit of the LR is performed.

The python package scikit-learn (Pedregosa et al., 2011) in version 0.18.1 was used to perform the training and testing. The same features, class-labels and CV folds as shown in Table 6.1 were used.

6.2.7 Classifier Performance Metrics

In order to quantify the performance of the RNN and to compare it to other classifiers the metrics recall, precision and F1 were computed. For these three metrics the performance from all three classes are reported. Additionally, the confusion matrix is computed showing the miss-assignment for the actual and predicted class. A more detailed description of these metrics is provided in Section 2.5.

6.2.8 Structural Model Assessment Metrics

The model quality of the snapshots is quantified by the two metrics GDTTS and Cα-RMSD, a comprehensive description can be found in Section 2.6.1.

6.3 Results

6.3.1 Overall Performance

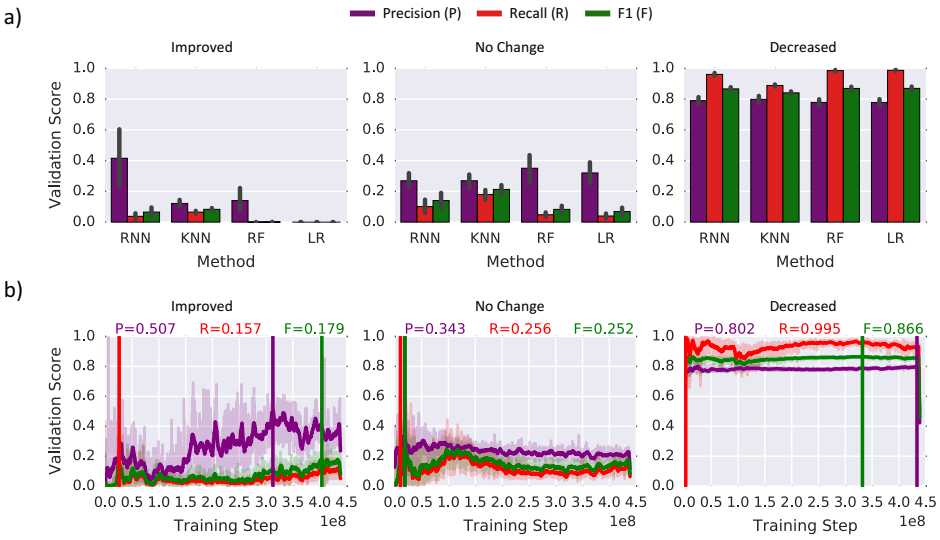

Figure 6.2: RNN performance comparison. a) Comparison of the recurrent neural network (RNN, with default parameter) against k-nearest neighbours (KNN), random forest (RF) and logistic regression (LR) on the full cross-validation set. The shown performance is measured on the validation set. b) Classification precision, recall and F1 as a function of training steps. The vertical lines indicate the best performance based on a moving average with a window-size of 30.

Figure 6.2a, left panel, shows that the spatial-temporal RNN model with default hyper-parameters is able to identify improvements in folds with a marked better precision than classical machine learning models. To be precise, the mean cross-validation precision for RNN, KNN, RF and LR have values of 0.415, 0.121, 0.139 and 0.000, respectively. For recall on the improved class the values are 0.037, 0.065, 0.001 and 0.000 for RNN, KNN, RF and LR, respectively. Values for the decreased class performance are similar for all models. The models RNN, KNN,

RF and LR produce a mean cross-validation precision of 0.790, 0.798, 0.778 and 0.777, respectively. The recall is 0.960, 0.888, 0.985 and 0.987 for RNN, KNN and RF and LR, respectively. All mean cross-validation values for precision, recall and F1 are available in Table 6.3 and a complete list of all results for each fold of the validation set is available in the supplemental material Table D.3.

Table 6.3: Mean CV performance. The mean performance for the three classes improved (I), no change (N) and decreased (D) for all 4 tested models is shown.

		Methods			
Class	**Metric**	**RNN**	**KNN**	**RF**	**LR**
I	F1	0.065	0.083	0.001	0.000
	Precision	0.415	0.121	0.140	0.000
	Recall	0.038	0.065	0.001	0.000
N	F1	0.140	0.213	0.083	0.069
	Precision	0.269	0.269	0.350	0.320
	Recall	0.101	0.179	0.048	0.040
D	F1	0.866	0.841	0.870	0.870
	Precision	0.790	0.799	0.779	0.778
	Recall	0.961	0.889	0.986	0.988

Figure 6.2b shows the validation score for precision, recall and F1 as a function of training steps for the three classes. The left panel for the improved class indicates that several million training steps are necessary to reach the best running average precision of 0.507. The best precision and recall for classes no change and decreased is reached early on during training (see panel centre and right). Furthermore, for these two classes the validation score stays stable during the 300 epoch training process.

The confusion matrix (CM) in Table 6.4 shows the miss-assignment of predicted classes versus the actual class for all 4 tested models for validation-fold 4 of the CV. For example, the RNN model predicted the correct true positive assignment for improved snapshots 1636 times and assigned the label improved incorrectly 314 times to no-change snapshots and 1653 times to decreased snapshots (see Table 6.4a). Compared to the three other models KNN, RF and LR this represents a notably better performance at identifying improved snapshots. For

Table 6.4: Confusion matrix. The four different sub-tables show CV (validation-fold 4) of the predicted and actual class assignment for improved (I), no change (N) and decreased (D) for a) RNN, b) KNN, c) RF and d) LR.

(a) RNN

		Predicted		
		I	**N**	**D**
Actual	**I**	1636	5497	14844
	N	314	3230	33327
	D	1653	6360	179923

(b) KNN

		Predicted		
		I	**N**	**D**
Actual	**I**	1556	3311	17110
	N	2386	6872	27613
	D	6690	12971	168415

(c) RF

		Predicted		
		I	**N**	**D**
Actual	**I**	34	614	21329
	N	25	2253	34593
	D	39	1644	186393

(d) LR

		Predicted		
		I	**N**	**D**
Actual	**I**	0	884	21093
	N	0	1526	35345
	D	0	1756	186320

KNN, seen in Table 6.4b, a similar number, i.e. 1556, of true positive improved snapshot assignments compared to RNN could be achieved. However, this comes with a large number of false positive assignments where the KNN incorrectly assigns the improved label to 2386 no-change snapshots and 6690 decreased snapshots. The RF model, see Table 6.4c, is hardly predicting the improved class. Here, 34 are true positive assignments and 25 and 39 are false positive assignments where the actual class is no-change and decreased, respectively. The LR model is not able to predict improved snapshots at all, i.e. the number of assignments is zero.

6.3.2 Markov Chain Interpretation of State Change

To assess how relevant the temporal component is in identifying improved, no change and decreased sections in the trajectory an analysis of the frequency of these states as a function of continuous segment length is performed. Figures 6.3a-c show their respective absolute frequency and their cumulative relative frequency. These histograms are heavily skewed to the right with a sharp peak at the minimum segment length. Looking at the frequencies of the segment lengths of the no change state in Figure 6.3b shows that these have shorter time spans compared to improved

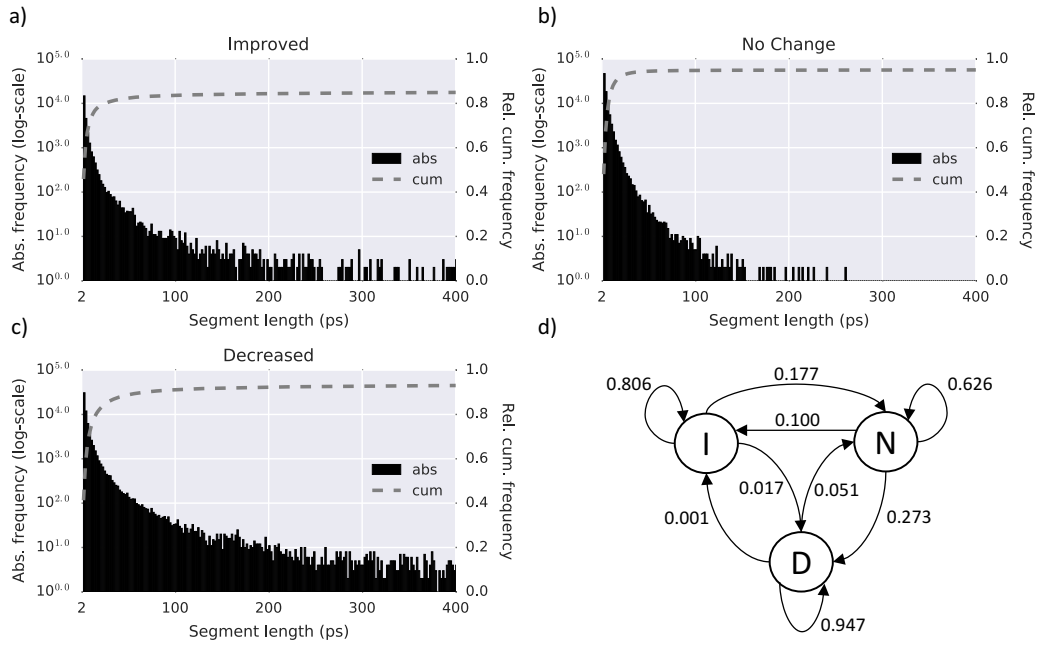

Figure 6.3: Segment length histogram. Frequency as a function of continuous segment length with a) improved quality, b) no change in quality and c) decreased quality. d) Markov chain model of the three states improved (I), no change (N) and decreased (D) visualised as circles and their directed transition probabilities shown as labelled arrows.

(Figure 6.3a) and decreased (Figure 6.3c) states. Additionally, the decreased state has a long tale of longer continuous segment lengths. Interpreting these results as a Markov-state model (Figure 6.3d) shows that indeed the states I (improved) and D (decreased) have a higher probability to reside in their states with 0.806 and 0.947, respectively, compared to N (no-change) with 0.626. Furthermore, when in state D the probability is low for transition back to I and N with values 0.001 and 0.051, respectively. The most likely transition away from I is to N with a probability of 0.177. Interestingly, the probability for a direct transition from I to D is a order of magnitude lower with 0.017.

6.3.3 Influence of Hyper-parameter Choice on RNN Performance

The hyper-parameter values of a RNN have great impact on its prediction performance (Pascanu et al., 2012), to this end a systematic exploration away from their initial default values was performed. Figure 6.4 shows the validation

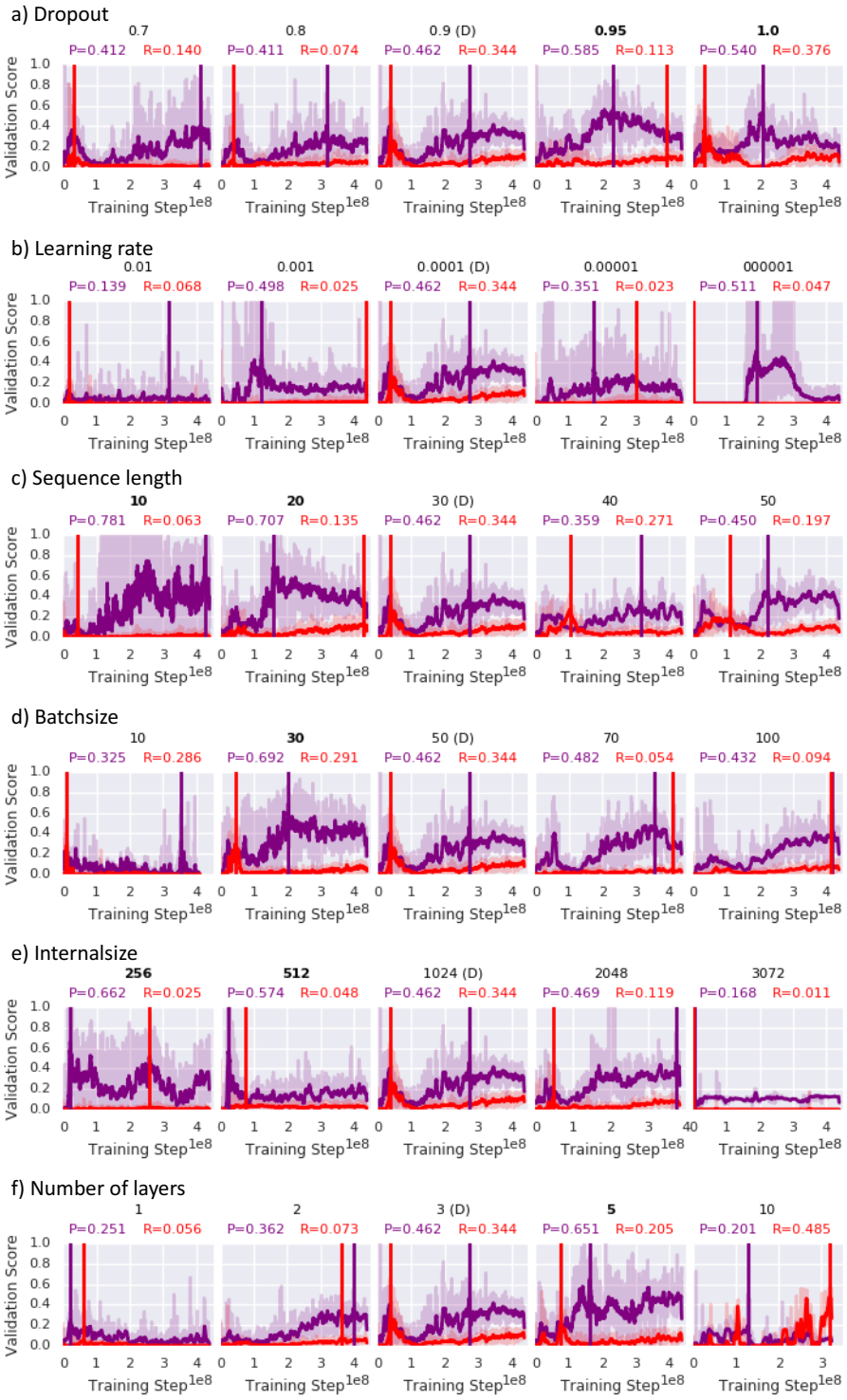

Figure 6.4: Exploration of RNN hyper-parameters (improved). The validation performance for recall (red) and precision (purple) of the improved class as a function of training steps for a) dropout, b) learning rate, c) sequence length d) batchsize, e) number of layers. The plot-title shows the tested parameter value where D in brackets indicates the default parameter and a bold value indicates > 0.05 improvement in precision over the default parameter. The vertical lines in each sub-plot indicate the best running average with a window size of 30, scores are shown in the sub-plot header.

score performance for the two metrics precision and recall of the improved class as a function of training steps on fold 4 of the validation set. The analysis focuses mainly on precision changes on the improved class as optimizations here would have the most immediate gain for model building. However, for completeness the training curves for no-change and decreased classes are shown in the supplemental material Figures D.1 and D.2.

The dropout parameter-values 0.95 and 1.0 have a marked improvement over the default parameter 0.9 where the recall could improve from 0.462 to 0.585 and 0.540, respectively. The results also show that too low values, i.e. 0.7 and 0.8 have a negative effect on precision performance. Changes to the learning rate from the default value 0.0001 have a negative impact on precision and recall. Too fast a learning rate, i.e. 0.01 and 0.001, lead to low recall scores, 0.068 and 0.025, respectively. The same is observed for the two slower learning rates 0.00001 and 0.00001 where the recall drops to 0.023 and 0.047. For precision, drops across all other tested learning rates are also observed. The parameter tests for sequence length show that shorter lengths produce the largest improvements on precision across all hyper-parameters. For lengths 10 and 20, values of 0.781 and 0.707 were achieved from an initial precision of 0.462 with the default length of 30. Longer lengths, such as 40 and 50 could not achieve an improvement over the default parameter value. Also, a change of the parameter batch-size from the default value 50 to 30 had a stark positive effect on the precision which increased to 0.692. Other values produced a similar performance to the default, i.e. 70 and 100, or decreased the precision, i.e. 10. The value 256 for the internal size of a GRU cell also leads to an increased precision with 0.662. However, this comes with a reduction in recall which drops to 0.025. Finally, testing for the number of layers in the network showed that, indeed, a deep neural network architecture is necessary to achieve good predictive power for the increased class. A RNN with 1 layer produced a decreased precision and recall of 0.251 and 0.056, respectively. With increasing layers, 2, 3 and 5, the precision could be improved to 0.362, 0.462 and 0.651, respectivley. However, too many layers, such as the tested 10 layer RNN, have a deteriorative

effect on precision. Here a drop to 0.201 in the experiment was observed.

6.4 Discussion

6.4.1 The Temporal Model is Successful at Identifying Improved Regions with Higher Precision

The results in Figure 6.2a show that the proposed RNN model with GRU cells is able to outperform classical machine learning methods such as RF, LR and KNN which are representative for state of the art classical machine learning algorithms and have been successfully applied to other bioinformatic domains (Pfeiffenberger et al., 2017; Liao and Chin, 2007; Li et al., 2004). In particular, the model presented here, based on default hyper-parameters, achieves a mean precision of 0.415 on the validation set of the CV compared to 0.121, 0.140 and 0.000 for KNN, RF and LR, respectively. The results also show that this precision could be further improved with optimized hyper-parameter choice. For example, the sequence lengths 10 and 20 produced a precision of 0.781 and 0.707, respectively, on the CV set 4 (see Figure 6.4c). This suggest that learned temporal dependencies of the used energy terms and distance metrics as input features are important to identify sections of fold improvements in the trajectory. This claim is further supported by looking at the transition probabilities between I, N and D in Figure 6.3d that are far from random. The transition probabilities of staying in their respective states from time-step t to t+1 is 0.806, 0.626 and 0.947 for I, N, D respectively. Which suggests that this sequential state awareness is important for predictive models on this particular classification problem. Furthermore, the combination explored in this thesis, where a time dependent sampling method is combined with a temporal RNN model that learns long and short term patterns from input features to identify improved conformations of proteins is promising over other methodologies. For example, sampling approaches based on Monte-Carlo simulations perform random sampling where the configuration of the system is not dependent on the last step (Caflisch, 1998), hence the information from the fourth dimension, i.e. temporal

component of protein folding can not be exploited as to where improvements are sampled.

6.4.2 The Balance Between Precision and Recall

The loss function L, defined as the weighted cross-entropy in Equation 6.7 with weights $\boldsymbol{\omega} = [0.05, 1, 10]$ led to the desired training success. Where a high precision for the improved class, and high recall for the decreased class was attained. The importance of a high recall of the decreased class is to act as a good filter to remove as many putative decreased quality snapshots as possible from the trajectory data. Conversely, the completeness of discovering all improved quality snapshots in a trajectory is not important here. High precision, i.e. lowering the false positive rate, is more important for correct model building from an ensemble.

An analysis of the observed recall/precision training curves on the improved class shows the delicate balance between these two. For this class, high recall values have the effect of a drop in precision and vice versa. The validation score curves as a function of training steps shown for sequence length 50 in Figure 6.4c visualize this phenomenon. Also, parameter which acquire high precision values during training have lower recall curves. For example, this is visible in Figure 6.4c for sequence lengths 10 and 20. The sequence length 10 and 20 produce a peak precision of 0.781 and 0.707, respectively. In the case of sequence length 10 the RNN model fails to exceed the peak recall of 0.063 and stays flat throughout the training. For sequence length 20, which has a lower precision than sequence length 10, a marked higher peak recall with 0.35 could be produced.

6.4.3 Model Complexity is Limited by the Amount of Data

Due to the high parameter space of deep neural networks, training of these usually requires large amounts of data (Najafabadi et al., 2015). The used data-set for training and testing contains ≈ 1.7 million snapshots. However, Table 6.1 shows that the percentage of improved quality snapshots in each validation fold in the CV ranges from 7 to 12 percent. The majority of snapshots are of decreased quality

making up 71 to 81 percent in each fold. An analysis of the validation score curves as a function of training steps for more complex models shows a drop in precision for the improved quality class. This is the case for parameters sequence length, internal size and number of layers. For sequence length, a marked drop in precision for the improved class can be observed for the tested lengths 30, 40 and 50. Similarly, for the internal size of h in a GRU cell, a too high value has a notable negative effect on precision. A size of 3072 lowers the peak precision to 0.168. Lastly, a high number of layers, i.e. 10, diminishes the precision gains compared to 5 layers (see Figure 6.4f). A way of improving the precision curves, and possibly train more complex networks, could be the inclusion of more refinement targets from, for example, CASP9 (Maccallum et al., 2011) and 10 (Nugent et al., 2014).

6.4.4 Future Directions

The analysis shows that the current RNN model is able to identify decreased and improved quality snapshots from the trajectory data. The performance with respect to increased quality snapshot is good with a mean precision of 0.415. One improvement to the model, which has not yet been implemented and tested, is the explicit use of the output y_t as an additional input to the feature vector at the next step x_{t+1} as shown in Figure 6.5. A use of this probability vector would explicitly model the putative current state, i.e. I, N, D. The Markov-chain interpretation of the trajectory data shown in Figure 6.3d makes it clear that there is a non-random transition probability between states. Thus, incorporating this knowledge explicitly to the model could lead to more accurate predictions. Furthermore, a deep RNN model proposed by (Cho et al., 2014) to learn phrase representations for statistical machine translation has shown that such an approach has great potential.

Chapter 5 has shown that the novel method for protein-protein complex refinement via a MD based sampling process is able to generate improved solutions. Similar to the RNN applied to trajectories of protein monomers, the proposed RNN model could be trained on the trajectories of protein dimers. The input features for these have to focus on the description of the molecular interactions

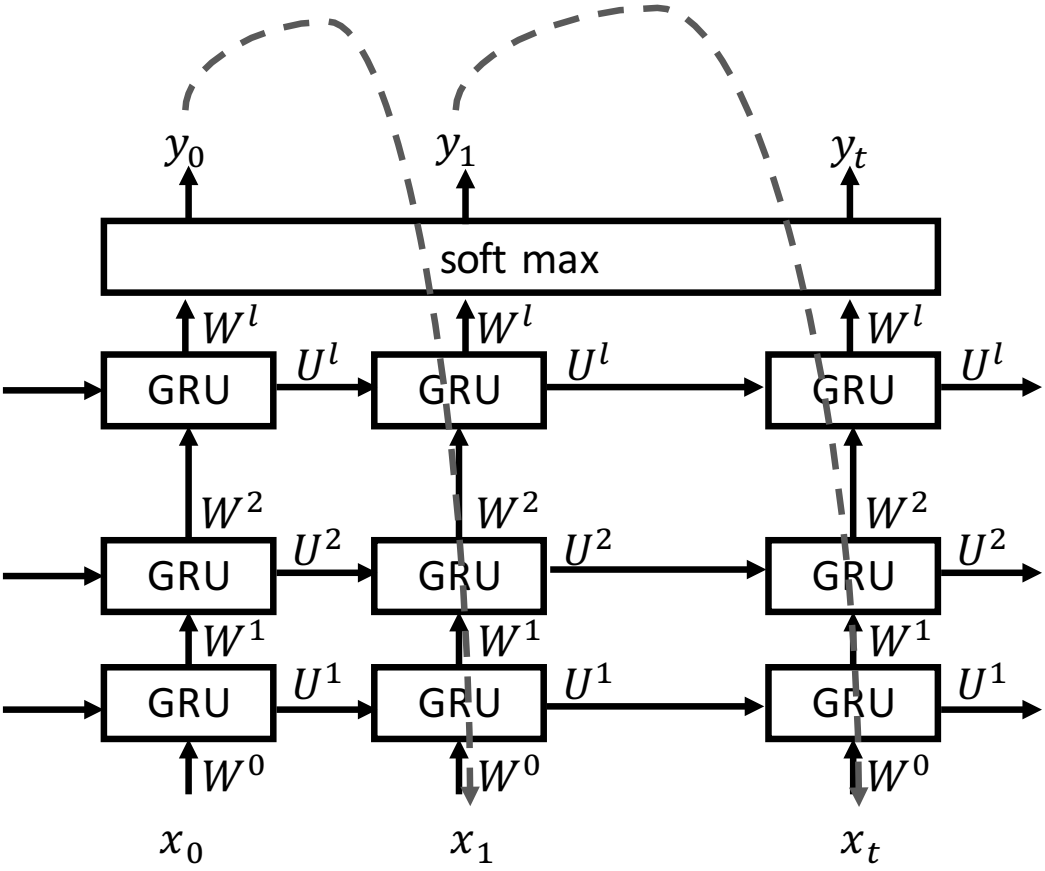

Figure 6.5: RNN model extension. The gray arrows indicate a propagation of the predicted output y_t to the next step x_{t+1} as an additional input.

of protein-interactions such as the functions provided in CCharPPI (Moal et al., 2015b).

CHAPTER 7

Understanding the Dynamics and Conformational Changes of Oncogenic RET-Kinase

7.1 Introduction

The RET-kinase is a member of the tyrosine kinase family (TK) where it plays an important role in kidney development and cell-linages that are derived from the neural crest (Arighi et al., 2005; Plaza-Menacho et al., 2006). The activation of RET is coupled to the binding of two glial cell line-derived neurotrophic factors (GDNF). These are known as GFL (GDNF family of ligands) and GFRα (GDNF family receptor alpha) where they bind to the extracellular domain of RET and bring together two entities of RET to facilitate auto-phosphorylation (autoP) (Lin et al., 1993; Angrist et al., 1998; Airaksinen et al., 1999). Work by Kawamoto et al. (2004) has identified the phosphorylation sites by mass spectrometric analysis, however, the temporal sequence of phosphorylation of these sites has long been illusive.

Recent work by Plaza-Menacho et al. (2014) identified early and late phosphorylation sites in RET and established the so called autoP trajectory not known before. The tyrosine residues Y1062 and Y687 are rapidly phosphorylated whereas phosphorylation of sites Y905, Y900, Y1015 and Y1029 occur at a

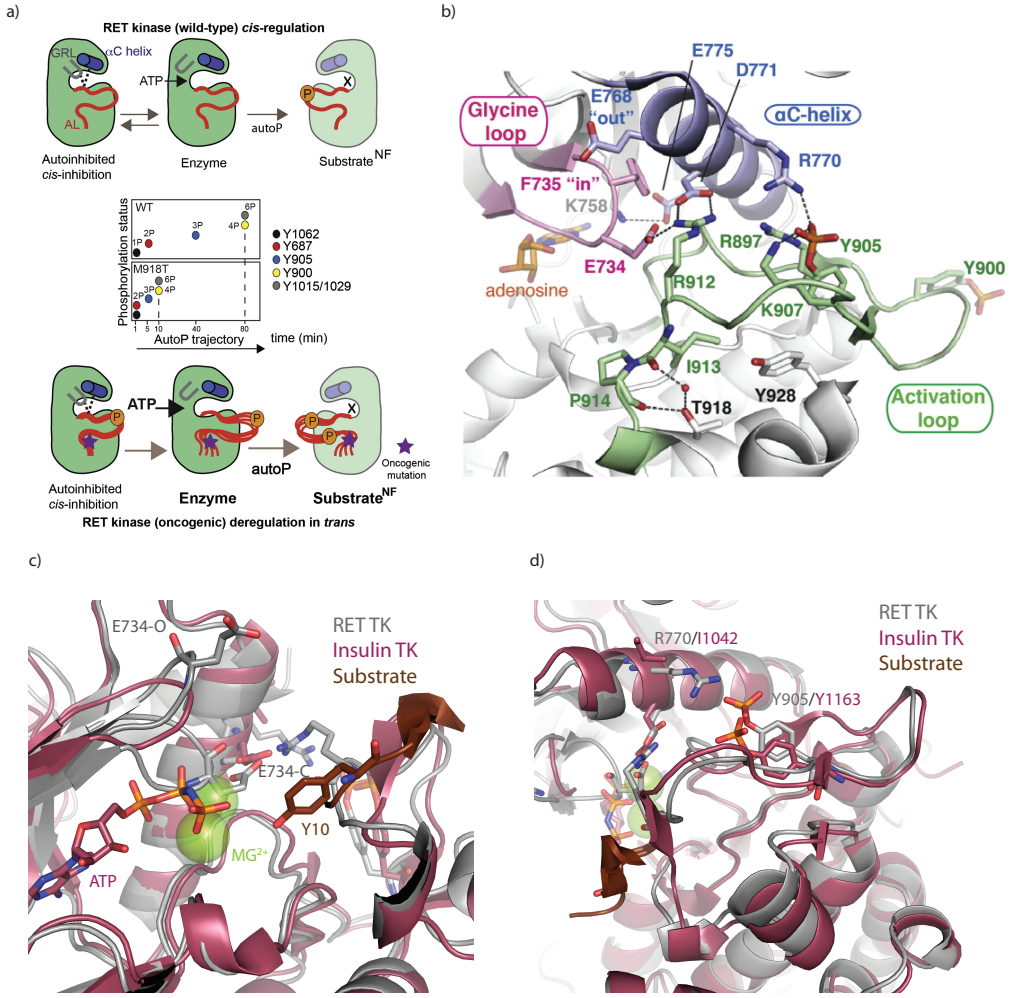

Figure 7.1: Structural building blocks and autoP in RET. a) Hypothesis of autoP deregulation in oncogenic RET (M918T). Top and bottom panel show the autoP mechanism and deregulation in wild-type and M918T, respectively. The center panel shows the enhanced autoP trajectory in M918T compared to wild-type, where late sites (Y905 and Y900) are more rapidly phosphorylated. b) Structural view of the RET catalytic domain with glycine rich loop (pink), αC-helix (blue) and activation loop (green). c) Close up view of the GRL site for RET (grey) and the insulin tyrosine kinase (purple) in its active state with bound substrate (brown) and ATP + Mg^{2+} (stick & sphere represetation). d) Close up view of the AL site for RET (grey) and the insulin tyrosine kinase (purple) in its active state with bound substrate (brown) and ATP + Mg^{2+} (stick & sphere represetation). Crystal structures shown in c & d are renderings of PDBs 4CKJ & 1IR3, respectivley. Figures a & b reproduced from Plaza-Menacho et al. (2014). Permission to reproduce this Figure a and b has been granted by Elsevier Inc.

later time. It was also found that the oncogenic mutation M918T, a main driver for multiple endocrine neoplasia type 2B (MEN2B), alters the autoP trajectory substantially. This mutation enhances the phosphorylation of the late sites dramatically and thereby these become early phosphorylation sites (Figure 7.1a). The mechanism how M918T enhances autoP is illusive and the authors propose the hypothesis that extended and more exposed activation loop (AL) conformations in M918T result in autoP increase for Y900 and Y905. However, the comparison of wild-type (WT) and mutant-type (MT) crystal structures show the same AL conformation, thus failing to provide structural evidence. Furthermore, the comparison of RET to insulin TK in its active state shows that RET AL is captured in the same state (Figure 7.1d). Interestingly, the crystal structure of RET has captured two states of the glycine rich loop (GRL), a regulatory element that controls the accessibility of adenosine triphosphate (ATP) to the active site, that are referred to as open and closed (Figure 7.1c). In the closed form the binding pocket is not able to accommodate ATP due to sterical clashes with the GRL as shown with an overlay of the structure of active insulin TK. This insulin TK structure was resolved with a substrate fragment and ATP coordinated to Mg^{2+} which could not be captured in RET. The GRL state of insulin TK structure shows an open-like conformation that is less extended when compared to RET structure due to the interaction with ATP. So far, it is not known for RET what the general GRL state preference is, even less so the interplay between GRL and AL and how M918T alters their behaviour causing accelerated autoP.

Motivated from this lack of understanding extensive i*n-silico* biophysical experiments are performed to illuminate the mechanism of autoP for Y900 and Y905 and its enhancement by M918T from a dynamic protein-motion perspective. To be precise, $\approx$ 5µs cumulative simulation-data is shown and analysed from standard molecular dynamics (MD) simulations, metadynamics (MetaD) simulations and directed pull-force experiments. The data suggest a revised view of RET-autoP function, and builds upon the initial work by Plaza-Menacho et al. (2014) where it is proposed that a cis auto-inhibitory contact between residues

E734-R912-D771 is crucial for regulation of autoP. In the following result sections it is shown that such a state is hardly formed in WT due to a predominantly open GRL. Furthermore, it is argued that such an open GRL state is necessary to acquire the so called AL ”in” conformation where sites Y900 and Y905 are less accessible and thus, become phosphorylated late in the autoP trajectory. From the analysis of the M918T trajectory data a conclusion can also be drawn as to how the autoP trajectory is enhanced. The data suggests a stabilisation of the P-pocket that promotes AL ”out” conformations that are further stabilized by a shift of the GRL state from predominantly open to an so called intermediate state.

7.2 Methods

7.2.1 Structure Preparation

The initial coordinates of the RET-kinase catalytic domain were obtained from the PDB entries 4CKJ and 4CKI for wild-type and the oncogenic mutation M918T, respectively. Missing side-chain atoms were modelled with SCWRL (Krivov et al., 2009) and the missing disordered segment from residue 822 to 844 with loopy (Xiang et al., 2002). The three other tested mutations , i.e. E734A, D771A and R912A, that are involved in the tether between the GRL and AL were introduced by in-silico mutagenesis with SCWRL.

7.2.2 Simulation Setup

Simulations were performed with GROMACS 4.6 (Hess et al., 2008) and the G54a7 force field (Schmid et al., 2011) with the Vienna-PTM 2.0 (Petrov et al., 2013) extension for force-field parameters of the phosphorylated Y905. The system was solvated with the explicit solvent model SPC/E (Chatterjee et al., 2008) and the charge was neutralized by Na^+ and Cl^- counter ions with a concentration of 0.15 mol/litre. The simulation box with periodic boundary conditions has a cubic shape with a 12 $\AA$ solvent-buffer between the bounds of the box and the protein. The

system was subject to a steepest decent energy minimization with a maximum of 50000 steps where the energy step size has a value of 0.01 and a pre-mature stopping condition if a force of < 1000 kJ mol^{-1} nm^{-1} is reached. Prior to production run the system was equilibrated in two steps. The first step consisted of a 100 ps long NVT simulation to increase the temperature of the system from 0 K to 300 K using V-rescale (Bussi et al., 2007). The second step of equilibration consisted of a 100 ps NPT simulation to increase the pressure of the system to 1 bar with Parrinello Rahman pressure coupling (Berendsen et al., 1984). Furthermore, during the two step equilibration position restraints of 1000 kJ mol^{-1} nm^{-1} were applied to all heavy atoms of the polypeptide chain (i.e. excluding solvent atoms).

The unrestrained production run was performed for 250 ns for each of the four conditions tested per mutation (i.e. wild-type (WT), M918T, E734A, D771A and R912A) which sums up to a cumulative simulation time of 1 µs for each mutation and 5 µs in total of unrestrained classical MD simulation data. A summary of these simulation set-ups can be found in Table 7.1. For all simulations, i.e. equilibration and production run, a leap-frog integrator with a Δt of 2 fs was used. Coordinates, velocities and energies and forces were saved every 2 ps. Furthermore, long range electrostatic interactions were treated with the Particle Mesh Ewald method (Darden et al., 1993) using a cutoff of 10 Å.

7.2.3 Metadynamics Simulation Setup

The metadynamics simulations with collective variables (CV) CV1 and CV2 were performed for WT and M918T with a closed GRL starting conformation and phosphorylated Y905 (see Table 7.1). The CV1 is defined such that

$$CV1(R) = 1/N \sum_{\gamma \in CM_{grl}} (D_\gamma(R) - D_\gamma(R_{ref}))^2. \tag{7.1}$$

This describes the formation of a set of atom to atom contacts, CM_{grl}, that are not observed in both, the closed and open GRL conformations. Thus, only unique

Table 7.1: RET simulation overview. Table summarises the simulation method (Simulation), mutation, occupation state of the active site (Pocket), conformation of the GRL loop at simulation start (GRL-Start), phosphorylation status of Y905, where P and NP indicate phosphorylated and unphosphorylated, respectively and simulation time.

Simulation	Mutation	Pocket	GRL-Start	Y905	Simulation Time
MD	WT	apo	open	P	250 ns
MD	WT	apo	closed	P	250 ns
MD	WT	apo	open	NP	250 ns
MD	WT	apo	closed	NP	250 ns
MD	M918T	apo	open	P	250 ns
MD	M918T	apo	closed	P	250 ns
MD	M918T	apo	open	NP	250 ns
MD	M918T	apo	closed	NP	250 ns
MD	E734A	apo	open	P	250 ns
MD	E734A	apo	closed	P	250 ns
MD	E734A	apo	open	NP	250 ns
MD	E734A	apo	closed	NP	250 ns
MD	D771A	apo	open	P	250 ns
MD	D771A	apo	closed	P	250 ns
MD	D771A	apo	open	NP	250 ns
MD	D771A	apo	closed	NP	250 ns
MD	R912A	apo	open	P	250 ns
MD	R912A	apo	closed	P	250 ns
MD	R912A	apo	open	NP	250 ns
MD	R912A	apo	closed	NP	250 ns
MetaD	WT	apo	closed	P	67 ns
MetaD	M918T	apo	closed	P	67 ns
Pull	WT	apo	closed	NP	10 ns
Pull	M918T	apo	closed	NP	10 ns

contacts are included in the set. For CV2, with definition

$$CV2(R) = 1/N \sum_{\gamma \in CM_{al}} (D_\gamma(R) - D_\gamma(R_{ref}))^2, \tag{7.2}$$

formalizes the atom to atom contacts of AL residues (CM_{AL}). Function $D_\gamma(R)$ defines a sigmoid distance function such that

$$D_\gamma(R) = \frac{1 - (r_{\gamma/r_\gamma^0})^n}{1 - (r_\gamma/r_\gamma^0)^m}, \tag{7.3}$$

where $n = 6$ and $m = 10$. A contact γ is defined as a distance < 5 Å between back-bone atoms N, Cα, O of two residues.

The preparation, energy minimization and equilibration is performed using the same protocol as described in Section 7.2.2. The production run with CV1 and CV2 was performed with PLUMED2 and GROMACS 4.6. The Gaussian addition is deposited every 2 ps with $\sigma = 1$, a bias factor of 10 and an initial height of 5 kJ mol^{-1}. The sampling was performed for 67 ns and snapshots were saved every 2 ps.

7.2.4 Pull Simulation Setup

The pull simulation measured the force magnitude required to extended the AL from an "in" to an "out" conformation with constant velocity. The coordinates for the "in" conformation of the AL was obtained from PDB 1XPD and modelled onto structure 4CKJ and 4CKI for WT and M918T, respectively. The vector for the direction of the pull movement was determined by the vector difference between position vectors of Cα-901 in "in" and "out" AL conformations. Prior to the production run of the pull-simulation, the system was energy minimized and equilibrated as described in Section 7.2.2. During the pull simulation all heavy atoms, except for atoms in the AL (residues 890 to 930), were position restrained with a force of 1000 kJ mol^{-1} nm^{-1}. The simulation extended the loop over a time of 10 ns and the force vector for dimensions x, y, z were recorded every 0.002 ps. From this the force magnitude was computed such that:

$$F = \sqrt{F_x^2 + F_y^2 + F_z^2}. \tag{7.4}$$

7.2.5 Metrics

The following metrics were computed from the trajectory data:

GRL z-axis: describes the open and closing movement of the GRL along the imaginary z-axis of Cα-E734. The reference closed and open conformation were obtained from the PDB 4CKJ. The centre point, referred to as the

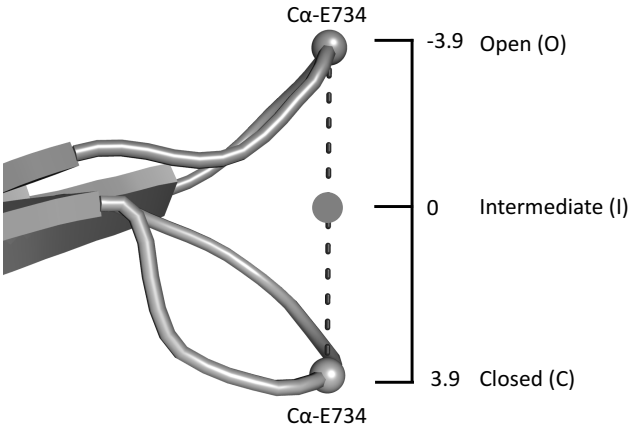

Figure 7.2: Explanation of the GRL z-axis. The GRL movement was quantified as the movement of the Cα-E734 along an imaginary z-axis. The open (O) and closed (C) states are observed from PDB 4CKJ. Their extension from open to closed ranges on the z-axis from -3.9 to 3.9. The self defined intermediate state (I) is the centre position between O and C and has the value 0 on the z-axis. All units are in $\mathring{A}$.

intermediate position, is defined as 0. A negative or positive deviation from this point is defined as a movement to an open or closed GRL conformation, respectively (see Figure 7.2).

AL RMSD: quantifies the conformational change around the P+1 pocket. The RMSD calculation included the Cα-atoms from residues 907-915. Each snapshot of the trajectory was first optimally superimposed to the reference crystal structure before RMSD calculation.

RMSF: The root mean square fluctuation (RMSF) is the average RMSD of an atom over time, and is defined such that:

$$RMSF = \sqrt{\frac{1}{N}\sum_{n=1}^{N}((x_n - x_0)^2 + (y_n - y_0)^2 + (z_n - z_0)^2)}, \qquad (7.5)$$

where N is the number of frames and x_0, y_0 and z_0 the starting coordinates of the atom in the trajectory. The RMSF was calculated for the Cα atom of each residue.

7.3 Results

7.3.1 Oncogenic RET induces Conformational Shift of GRL and AL

The analysis of the trajectory data from unbiased MD simulations show that WT and oncogenic RET sample two distinct GRL conformations (see Figure 7.3a, left). In the four different conditions tested, the WT samples predominantly open conformations in the 250 ns simulations, regardless of the initial starting GRL conformation. Only for condition O/M918T/Apo/P a shift to an intermediate to closed conformation is observed from simulation time 150 ns onwards. Simulations of the oncogenic mutant M918T show a stable intermediate to closed conformational state of the GRL in 3 out of 4 tested conditions. In simulation O/M918T/Apo/NP the GRL stays stationary in the open conformation. Nevertheless, the start of a transition from open to intermediate state at time-point 225 ns can be observed. The aggregated simulation data from all four conditions shown in Figure 7.3b summarizes the observed sampling. It shows that WT RET has a stable open GRL conformation with low variance from the first to the third quartile of the data. On the other hand the aggregated data for M918T shows a median at an intermediate conformational state with high variance from the first to third quartile.

The AL RMSD deviation around the P+1 pocket in the four tested conditions is shown in Figure 7.3a, right panels. The data shows that deviations of at least 4 Å are sampled regardless of WT or MT. Furthermore, M918T exhibits a more stable sampling as indicate by low fluctuations from one to the next time-point and no drift to other RMSD regions. The WT simulations exhibit more variation in AL RMSD an can markedly change during the time course of the 250 ns simulation. For example, trajectories C/M918T/Apo/NP and C/M918T/Apo/P show the described behaviour. The aggregation of all trajectory data is shown in the box-plot of Figure 7.3c, where a significant difference between WT and M918T AL RMSD is seen. Furthermore, the variance between the first and third quartile is higher in WT

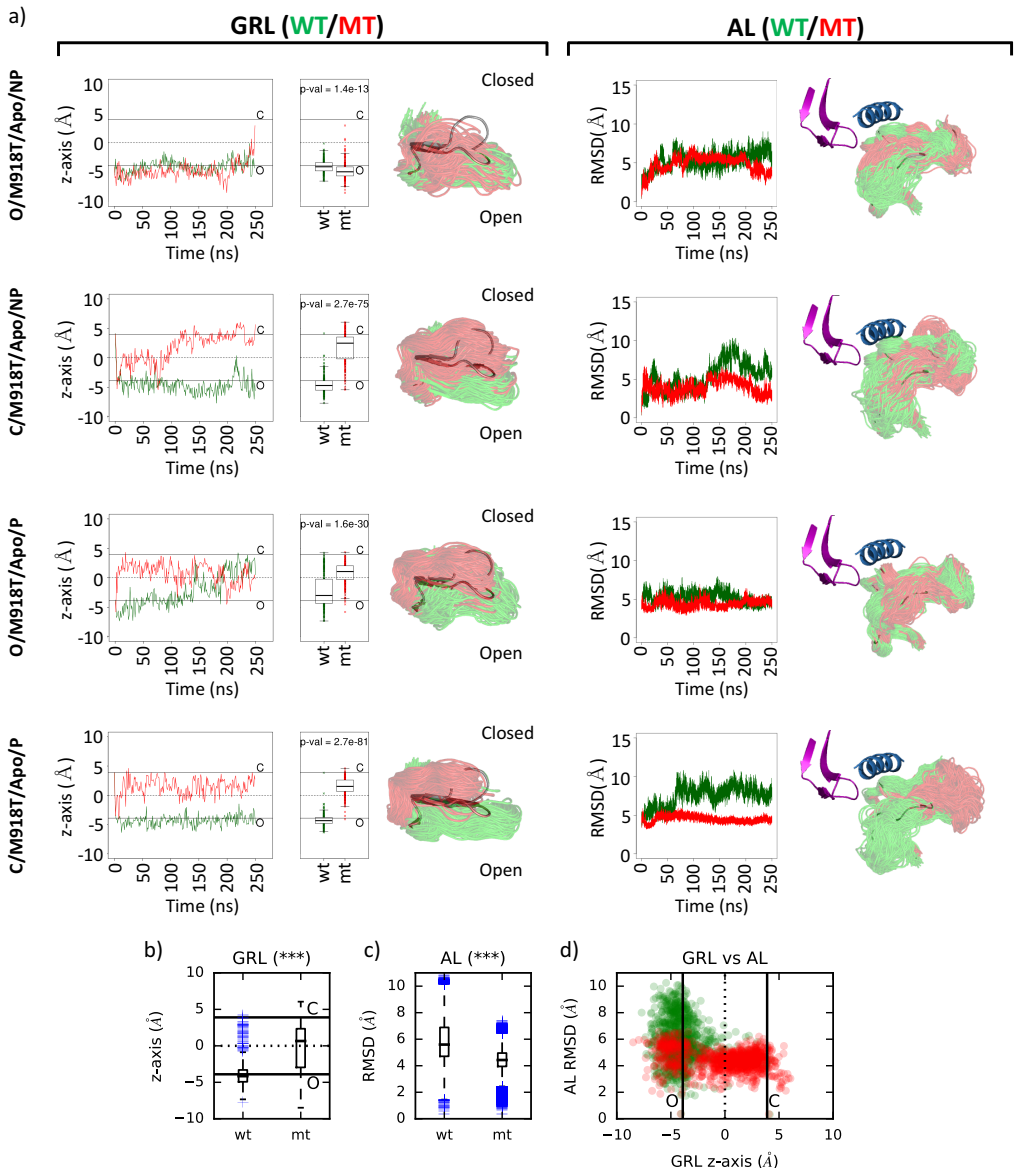

Figure 7.3: Dynamics and conformational states of GRL and AL in M918T. a) Left: change of GRL states as a function of simulation time for four different simulation conditions (rows). Right: change of AL conformation expressed as RMSD as a function of simulation time (rows). Data for wild-type (WT) and mutant types (MT) are shown in green and red, respectively. The 3D rendering shows the overlap of all sampled conformational states with a resolution of 1 ns, the gray rendering represent the conformations observed in the PDB 4CKJ. The aggregated data for all simulations for GRL and AL is shown in b) & c), respectively. The *** in brackets indicate significance with p-value < 0.001 between WT and MT data. d) 2D scatter plot visualizing the relationship between GRL conformation (x-axis) and AL conformation (y-axis) for WT (green) and MT (red).

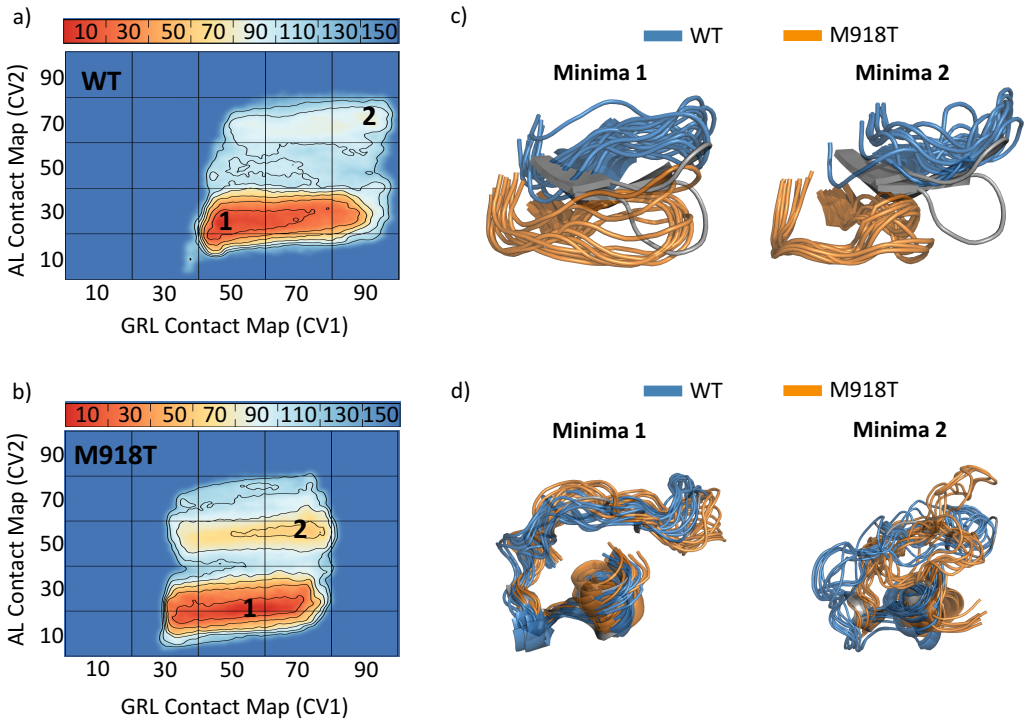

Figure 7.4: Free energy landscape of wild-type and M918T. a) & b) reconstruction of the free energy landscape of wild-type (WT) and M918T from metadynamic simulations with CV1 (GRL contact map space) and CV2 (AL contact map space). Shown energies are normalized to range from 0 to 160 for comparability. c) conformational clusters of GRL for minima 1 and 2 in WT (blue) and M918T (orange), gray rendering shows the conformations observed in PDB 4CKJ. d) conformational clusters of AL for minima 1 and 2 in WT (blue) and M918T (orange), gray rendering shows the conformations observed in PDB 4CKJ.

compared to M918T. The scatter plot in Figure 7.3d visualizes GRL movement versus AL RMSD. The data clearly shows that WT RET kinase sample a wide range of P+1 AL RMSD values and attain a stable open GRL conformation. Whereas the data for M918T shows the reverse effect. The P+1 AL RMSD is lower and more stable, whereas a large range of open and closed GRL conformations are sampled where most data-points fall within the intermediate to closed region.

7.3.2 Free Energy Landscape of WT and Oncogenic RET

The conformational free energy landscape was reconstructed from metadynamic simulations in order to collect further evidence of the distinct GRL and AL

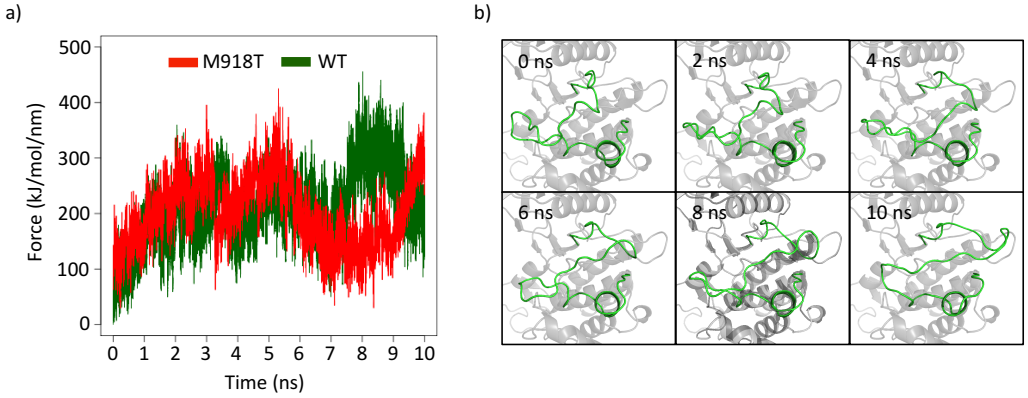

Figure 7.5: AL pull simulation of wilde-type and M918T. a) Force required to extent the AL from an 'in'-state to an 'out'-state over 10 ns. b) Visualisation of AL conformational states (green) during the pull simulation with a temporal resolution of 2 ns.

conformations in M918T. The results show a shift along CV1 (GRL CM) and CV2 (AL CM) in M918T (Figure 7.4b) when compared to WT (Figure 7.4a). The two analysed minima in both landscapes confirm the results from the unbiased MD simulations. For both minima the GRL of M918T has a distinct intermediate to closed conformation whereas for WT both minima capture an open GRL state (see Figure 7.4c). The AL conformations in minima 1 for WT and M918T are similar, however, in minima 2 mutant type M918T posses a distinct different AL conformation (see Figure 7.3d). The WT conformation of AL around the P+1 pocket notably diverges from the observed crystal structure conformation, seen in gray, as well as M918T, which exhibits less deviation from the crystal structure. Furthermore, the AL residue Y905 is distinctly more exposed and has a large deviation from both WT conformation and crystal structure.

7.3.3 A Force Perspective of Oncogenic AL Extension

The pull simulation with constant velocity, over a time-frame of 10 ns, shows that the force required to extent the AL from an "in" to an "out" conformation is similar up to 6 ns into the simulation, as seen in Figure 7.5a. Continuing from this point the force required to further extend the AL is markedly higher in WT than M918T.

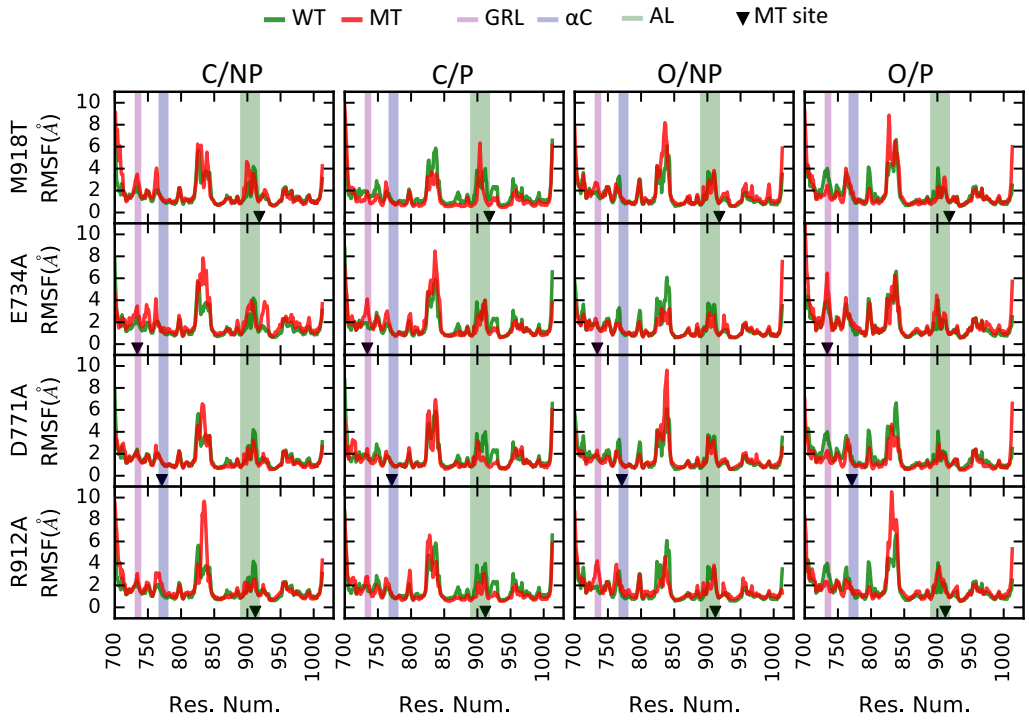

Figure 7.6: RMSF profile. Each row shows the RMSF profile between wild-type (WT) in green and mutant-type (MT) in red for the 4 mutations M918T, E734A, D771A and R912A. Each column represents the data from different simulation conditions where C and O stand for a open and closed starting conformation of the GRL and NP and P for an un-phosphorylated and phosphorylated Y905, respectively. The three coloured bars in each sub-plot indicate the location of the glycine rich loop (GRL), αC-helix (αC) and activation loop (AL). The black triangle marks the location of the respective mutation.

A visual inspection of the conformations at these time points shows that they are qualitatively similar to the conformation observed in minima 2 of the free energy landscape of WT (see Figure 7.4d).

7.3.4 GRL and AL Conformational States of E734A

The mutation E734A in the GRL causes a shift from a mainly closed to an intermediate conformational state. This is especially pronounced in experiments starting from a closed GRL conformation (see Figure 7.7a, right panel, rows C/E734A/Apo/NP and C/E734A/Apo/P). However, the trajectory data from O/E734A/Apo/NP did not show this state preference where E734A was exclusively sampling open GRL conformations during the entire 250 ns simulation. The

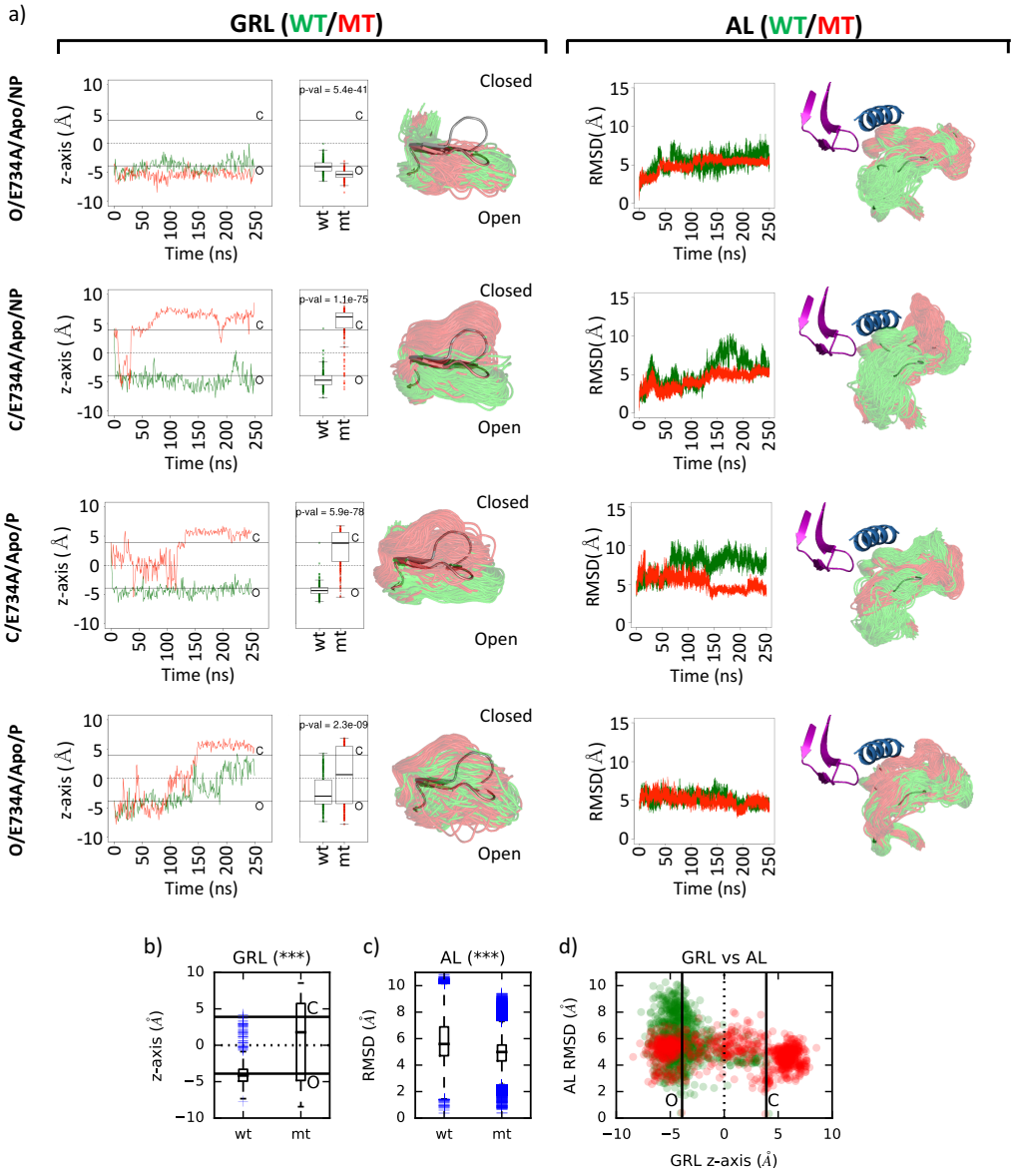

Figure 7.7: Dynamics and conformational states of GRL and AL in E734A. a) Left: change of GRL states as a function of simulation time for four different simulation conditions (rows). Right: change of AL conformation expressed as RMSD as a function of simulation time (rows). Data for wild-type (WT) and mutant types (MT) are shown in green and red, respectively. The 3D rendering shows the overlap of all sampled conformational states with a resolution of 1 ns, the gray rendering represent the conformations observed in the PDB 4CKJ. The aggregated data for all simulations for GRL and AL is shown b) & c), respectively. The *** in brackets indicate significance with p-value < 0.001 between WT and MT data. d) 2D scatter plot visualizing the relationship between GRL conformation (x-axis) and AL conformation (y-axis) for WT (green) and MT (red).

second trajectory starting from an open conformation, O/E734A/Apo/P, did show a transition to closed after 100 ns. The aggregated GRL data from all four trajectories shown in boxplot 7.7 summarizes the described observation from the single trajectory data. GRL states in E734A have a large variance from the first to third quantile with a median value at the intermediate to closed state.

The P+1 AL RMSD in E734A is distinct from WT in two trajectories: C/E734A/Apo/NP and C/E734A/Apo/P. The other two trajectories exhibit a similar RMSD when compared to WT. The aggregated AL data from all trajectories in Figure 7.7c shows that the median of E734A is lower than WT, however, the plot also shows that a number of outliers are observed with RMSD values ranging from 7-9 Å RMSD. Figure 7.7d, with AL RMSD on the y-axis and GRL change on the x-axis, visualises the relationship between these two. E734A has high occupation of closed and open GRL conformations. Along this axis of GRL conformations no change of AL RMSD is observed and does not reach the WT observed deviations.

7.3.5 GRL and AL Conformational States of D771A

The mutation D771A, located in the αC-helix, has no effect on GRL conformational states. WT and D771A favour a stable open conformation in all four tested conditions (see Figure 7.8a, left). The aggregation of all data as shown in the barplot in Figure 7.8b supports this, where the median for WT and D771A are both in an open configuration.

The AL P+1 pocket has a marked different RMSD compared to WT for conditions without phosphorylated Y905 as seen in Figure 7.8a, right. For the two other conditions similar RMSD trajectories are observed. However, the barplot of aggregated AL RMSD data, see Figure 7.8c shows similar median RMSD values. Finally, the scatter-plot in Figure 7.8d, shows that AL-RMSD (y-axis) and GRL state (x-axis), have a similar sampling of states during the 250 ns simulations between WT and D771A.

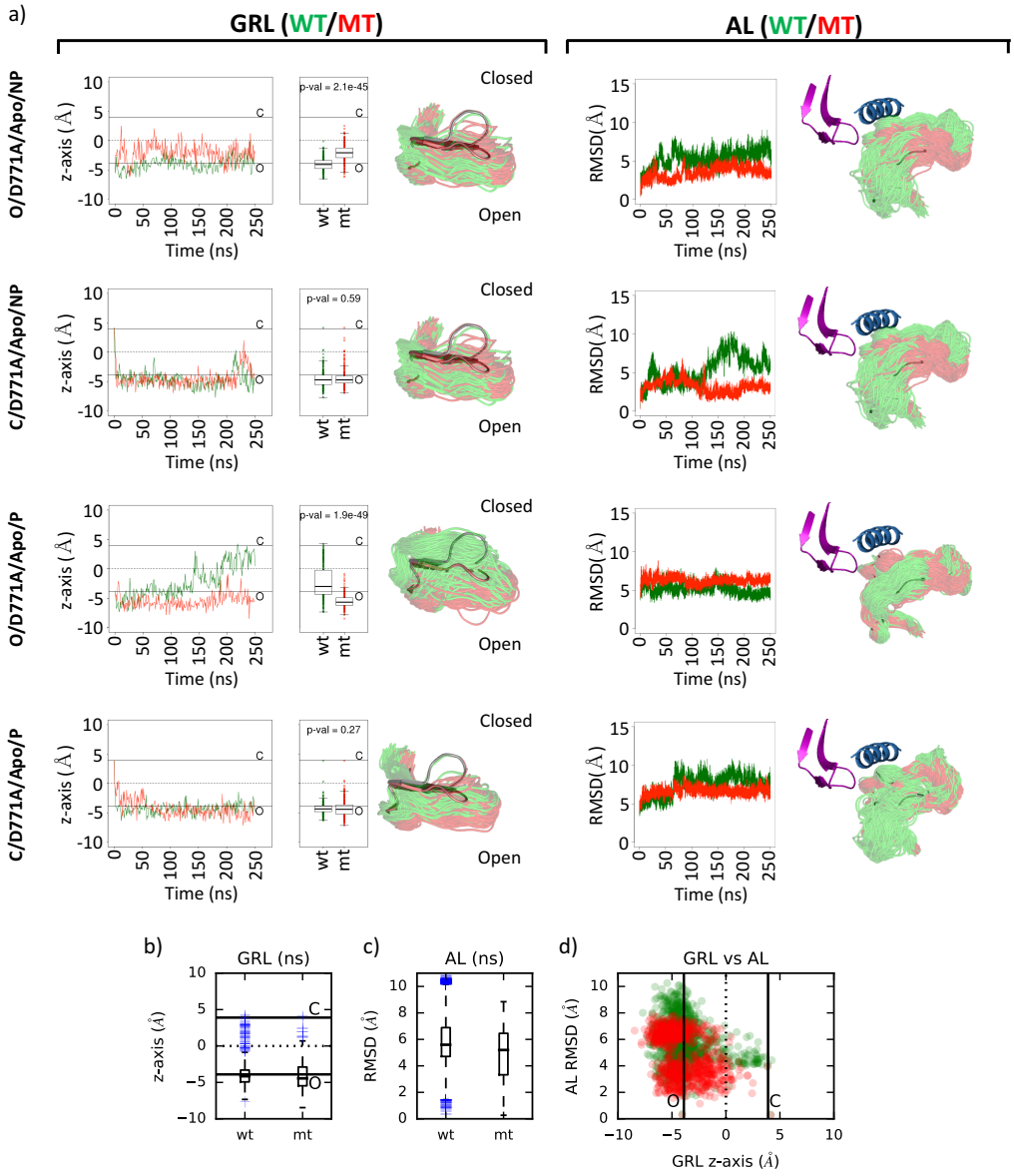

Figure 7.8: Dynamics and conformational states of GRL and AL in D771A. a) Left: change of GRL states as a function of simulation time for four different simulation conditions (rows). Right: change of AL conformation expressed as RMSD as a function of simulation time (rows). Data for wild-type (WT) and mutant types (MT) are shown in green and red, respectively. The 3D rendering shows the overlap of all sampled conformational states with a resolution of 1 ns, the gray rendering represent the conformations observed in the PDB 4CKJ. The aggregated data for all simulations for GRL and AL is shown b) & c), respectively. The *** in brackets indicate significance with p-value ¡ 0.001 between WT and MT data. d) 2D scatter plot visualizing the relationship between GRL conformation (x-axis) and AL conformation (y-axis) for WT (green) and MT (red).

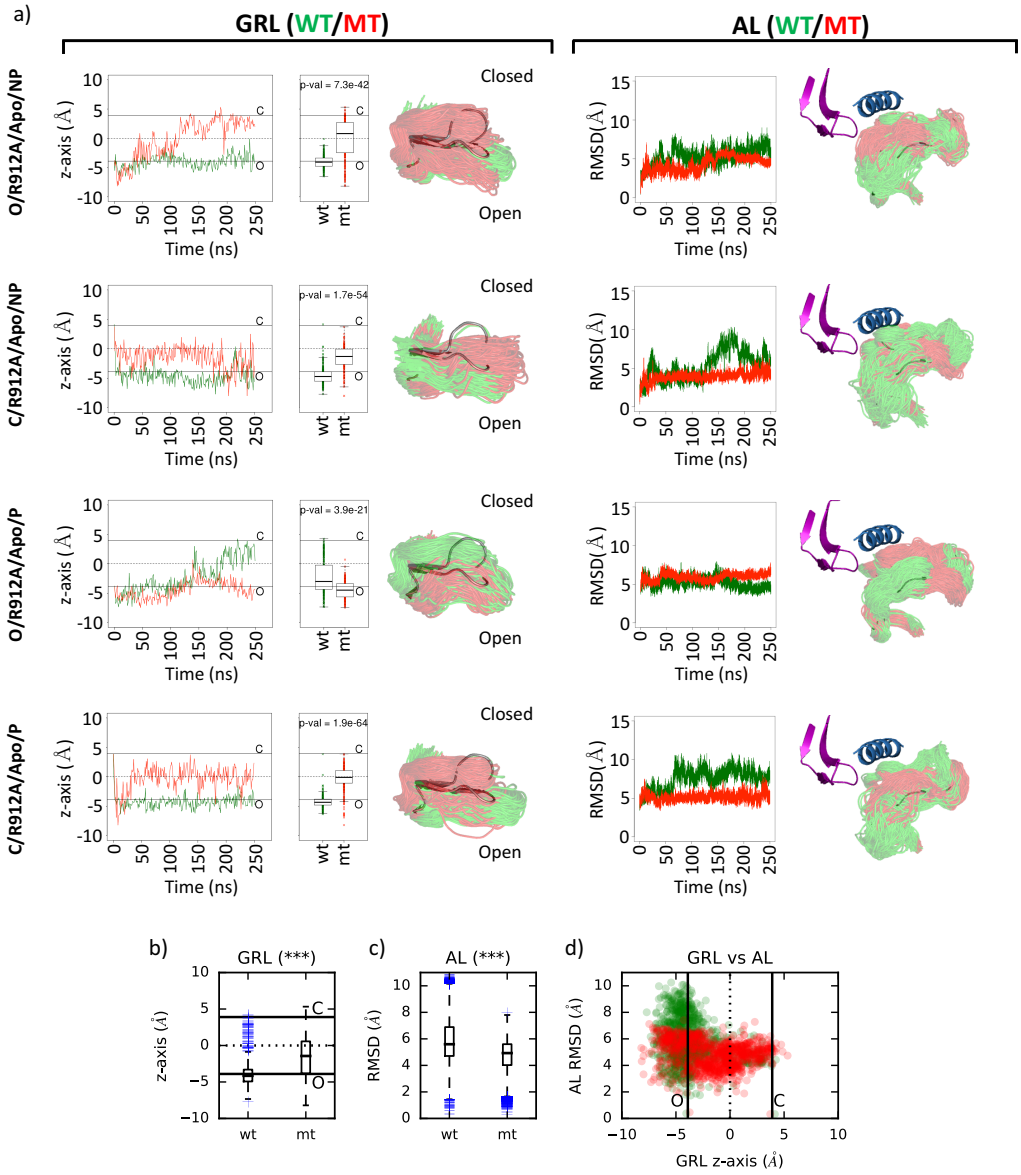

Figure 7.9: Dynamics and conformational states of GRL and AL in R912A. a) Left: change of GRL states as a function of simulation time for four different simulation conditions (rows). Right: change of AL conformation expressed as RMSD as a function of simulation time (rows). Data for wild-type (WT) and mutant types (MT) are shown in green and red, respectively. The 3D rendering shows the overlap of all sampled conformational states with a resolution of 1 ns, the gray rendering represent the conformations observed in the PDB 4CKJ. The aggregated data for all simulations for GRL and AL is shown b) & c), respectively. The *** in brackets indicate significance with p-value ¡ 0.001 between WT and MT data. d) 2D scatter plot visualizing the relationship between GRL conformation (x-axis) and AL conformation (y-axis) for WT (green) and MT (red).

7.3.6 GRL and AL Conformational States of R912A

The AL mutation R912A causes a conformational shift of the GRL conformation (see Figure 7.9a, left). In two out of four trajectories a stable intermediate GRL conformation is observed as seen for rows C/R912A/Apo/NP and C/R912A/Apo/P. In one trajectory, O/R912A/Apo/NP, a transition over 120 ns from open to closed is observed. The trajectory O/R912A/Apo/P showed a stable open conformation during the complete simulation. The aggregation of all GRL data for all trajectories (see Figure 7.9b) shows that most data points lie between an open and intermediate conformation, as indicated by the first to third quartile box.

Figure 7.9a, right, shows that the R912A AL RMSD is stable with a 5 Å deviation from the crystal structure for all 4 trajectories. In comparison, the WT AL RMSDs measure a much more dynamic P+1 site as indicated by transitions to different RMSD values. This is also reflected in the barplot for the aggregated AL RMSD data (Figure 7.9c), where the expansion of the box plot is higher in WT. The scatter-plot in Figure 7.9d also visualizes this. Here the red point cloud for R912A samples a wide range of GRL states, whereas in WT a wider range of different AL RMSD values are observed that are not sampled in R912A.

7.4 Discussion

7.4.1 WT Samples Predominantly Open GRL Conformations and Prefers an ”In” AL Loop Conformational State

The data suggest that WT GRL prefers an open GRL conformation. The four different simulation conditions all produced a trajectory of a stable open conformation. This is further supported by the reconstruction of the conformational free energy landscape from metadynamic simulations. The two analysed minima both showed an open conformation (Figure 7.4a and c). Furthermore, the evidence for the cis-inhibitory mechanism, proposed by Plaza-Menacho et al. (2014) and seen in Figure 7.1b, is weak for WT RET due to an open GRL. The only exception

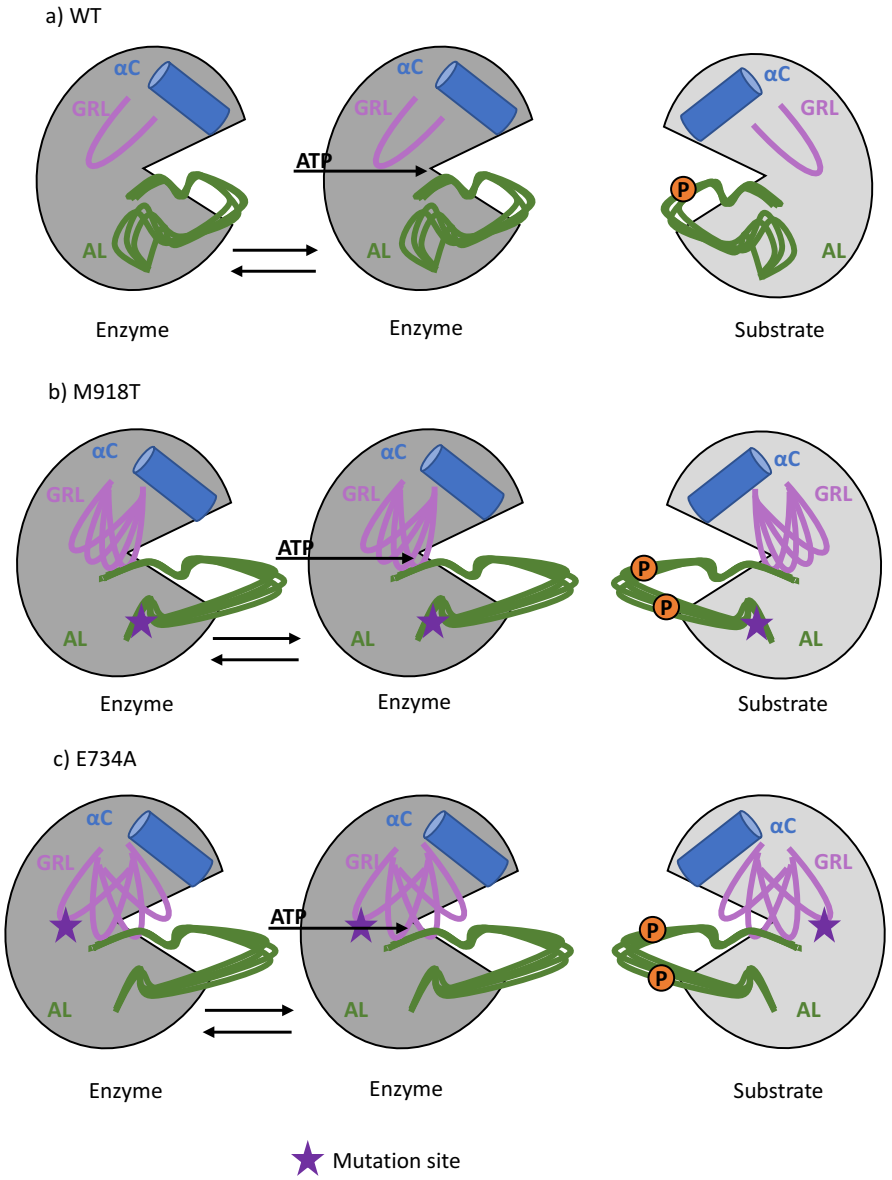

Figure 7.10: Revised RET function. a) WT RET samples predominantly open GRL conformations which allow for an AL "in" conformation where late phosphorylation sites Y900 and Y905 are less accessible. Furthermore, the open GRL state leads to a higher disassociation of ATP, thus lowering enzymatic activity. b) Oncogenic M918T leads to a stabilisation of the AL "out" conformation with the result that the GRL samples increased intermediate states that further stabilize AL "out" where sites Y900 and Y905 become more accessible. Additionally, this shift of GRL conformations leads to a lower disassociation rate of ATP thus allowing for better enzymatic activity. c) E734A deregulates GRL which now samples a wide range of open and closed conformations. This dynamic behaviour of GRL inhibits the AL "in" conformation, hence, resulting in more exposed Y900 and Y905.

was one simulation where a slow transition from open to closed at time 150 ns to 250 ns is observed. Thus, it may be concluded if such a tether is present in WT RET, it would only represent a weak transient state.

In addition, the data suggest that the site from residues 907 to 915 samples a wide range of different conformations. The enhanced sampling in this region leads to an ”in” conformation of the AL where the two late phosphorylation sites, Y900 and Y905, are less accessible. This ”in” conformation is also supported by the meta-dynamic simulations. The second minima from the reconstructed conformational free energy landscape has qualitatively a similar conformation (Figure 7.4d). An illustration of the described mechanism is shown in Figure 7.10a.

7.4.2 M918T Induces Conformational Shifts to Intermediate GRL and ”out” AL Conformations

The oncogenic mutation M918T resulted in a shift of the GRL from a stable open to a more flexible GRL where the median conformational state can be described as intermediate to closed (Figure 7.3). This finding is also supported by the two analysed energy minima from the reconstructed free energy landscape, where in both cases closed GRL conformations are observed (Figure 7.4c). From this an explanation can be drawn as to the lower disassociation rate for ATP in M918T compared to WT, where $K_d = 1.1 \pm 0.9$ and $K_d = 14.3 \pm 0.3$, respectively (Plaza-Menacho et al., 2014). The hypothesis is that the intermediate GRL conformation allows for more stable binding of ATP and thus less disassociation events with the effect that phosphorylation of the substrate is enhanced.

Furthermore, three points of evidence could be collected that support the hypothesis of an exposed AL and hence, that M918T RET acts as a better substrate in autoP. i) The data shown for AL conformations in Figure 7.3a indicates that more accessible conformational states around the phosphorylation sites Y900 and Y905 are sampled, ii) the second minima of the free energy landscape shows an exposed AL conformation (Figure 7.4) and iii) the measured force for an AL extension is lower in M918T.

From this data, a mechanism is concluded that mutation M918T prevents the sampling of AL ”in” conformations by stabilizing the region around the P-pocket. Thus, allowing for an intermediate to closed GRL with transitions to the cis-state, which is further stabilizing the sampling of more exposed AL ”out” conformations. A schema illustrating this mechanism is shown in Figure 7.10b.

7.4.3 E734A Causes Deregulation of GRL

The other mutation that causes an enhanced autoP trajectory is E734A. The analysis of the GRL movements from the trajectory data showed that a wide range of open and closed conformations are sampled. This is also reflected in the RMSF profile where the GRL segment has a marked higher fluctuation (Figure 7.6). Given the evidence, this allows for the conclusion that the loss of the interaction with R912 leads to a deregulated movement of GRL. This also suggest that the frequent sampling of closed GRL conformations prevents the sampling of AL ”in” conformations thus explaining the enhanced phosphorlation of the late sites as measured by Plaza-Menacho et al. (2014). An illustration of this mechanism is shown in Figure 7.10c.

7.4.4 GRL State is Coupled to AL Extension

From the data presented and the discussion of the results and their mechanism a conclusion can also be drawn on the relationship between GRL state and AL extension. Figure 7.11 shows the proposed model where the GRL state is coupled to AL extension. Here, it is proposed that only open GRL conformations allow for sampling of the AL ”in” conformations and that this sampling is also dependent on the duration the GRL open state. The trajectory data of WT simulations C/E734A/Apo/NP and C/E734A/Apo/P show that higher RMSD values are only reached after 50-100 ns. Thus, transitions to this state take time. Additionally, this transition to AL ”in” can only happen when it is not disrupted by closed GRL conformations. This mechanism is supported by trajectory O/E734A/Apo/P where a transition from open to closed GRL starts from 100 ns with the result

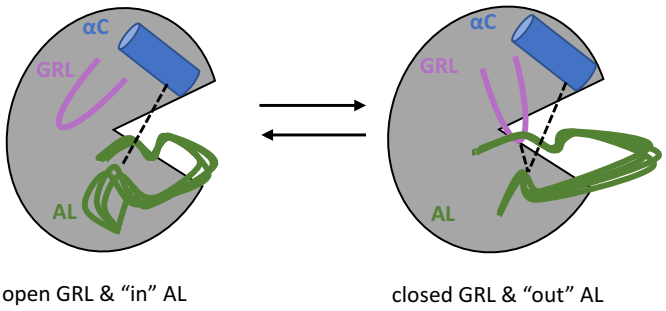

Figure 7.11: Dependency of GRL state and AL extension. A proposed model how GRL state is responsible for accessibility of Y900 and Y905 in the AL. On the right site a continuous open GRL is shown which is a prerequisite for sampling AL ”in” conformations. When the GRL closes, seen on the right, it forms the cis inhibitory tether with E734-R912-D771 which promots and stabilises the AL ”out” conformation, leading to a better accessibility of Y900 and Y905.

that AL ”in” conformations are prevented. Thus, the deregulation of GRL movement by mutation E734A, sampling open and closed conformations, inhibits the slow transition to AL ”in” and favours AL ”out” with the result of enhanced phosphorylation of Y900 and Y905.

7.4.5 Future Directions

The shown reconstruction of the free energy landscapes in Figure 7.4a and b are the result of a 67 ns long metadynamics simulations. These landscapse have not yet converged and could change when extended. The convergence of such a landscape can take 700 ns as shown in a study by Sutto and Gervasio (2013).

Additionally, further metrics to measure different aspects such as the the cis-inhibitory state in more detail could help to clarify questions. For example, currently it has not been quantified how exactly the interaction as a function of time between residues R912-E734-D771 is altered in WT and M918T. Also, it is still not fully understood how the mutation M918T stabilises the AL loop from residue 905 to 917. A measurement of all residue-residue contacts with M918 and T918 and an analysis of their differences could help to illuminate this mechanism.

CHAPTER 8

Epilogue

In this thesis, new methods have been developed to i) improve the identification of near native binding sites of protein-protein complexes, ii) improve the model quality of predicted protein folds, iii) improve the model accuracy of docked protein-protein complexes and iv) identify regions of improved protein monomer model quality from MD trajectories. Furthermore, an *in-silico* study was conducted to understand RET-kinase auto-phosphorylation in wild-type and oncogenic mutant-type M918T.

The recurring theme of this thesis is the attempt to predict and understand the structure and function of protein systems from a dynamic perspective. The ranking method for protein-protein docking presented in Chapter 3 approaches the identification of correct binding sites as a dynamic process where an ensemble of encounter states substantially contribute to the binding specificity. Integrating this information into the ranking method was my aim and represents a deviation from the classical way of ranking, where docked solutions are ranked based on a scoring function that evaluates the energy in isolation of other solutions. A side-effect that emerged from the analysis of the produced rankings was that often a funnel of attraction was observed. Where surrounding clusters that are close to the near-native binding-site cluster were ranked highly too. This was not much discussed or systematically quantified, and at the moment is just a hypothesis. However, an systematic analysis whether such an funnel of attraction is prevalent across a wide range of different protein complexes will be interesting as well as

whether this information can be further exploited to improve the identification of native binding sites.

The optimization methods for predicted protein folds and complexes presented in Chapter 4 and 5 attempted to produce higher quality models by simulation. During the simulation, the surface of the protein energy landscape is naturally explored by following the negative gradient vector of the potential energy function. The results showed that such a sampling can successfully produce more native like states of protein folds and complexes. However, the identification of these with classical scoring functions less so. This was addressed by exploring the question whether the patterns of energy in time can give more information whether states are reached that resemble more native like conformations. And if so, can these patterns be learned to reliably predict the state changes from more native like conformations to less native like conformations and *vice versa*? The results presented in Chapter 6 to that question are promising. The temporal deep RNN model that was trained from the trajectory data had significantly better performance than other machine learning models that do not consider temporal dependencies.

Subsequently, now that it was shown that a neural network model can be trained to identify transition to states that are closer to the native state of a system, the question could be asked: is it possible to define a model that learns how to fold a protein on its own? The rapid progress in deep learning seems to suggest that such considerations are no longer pipe dreams. For example, the latest success in reinforcement learning demonstrated that a neural network could learn to play the complex Chinese board game Go from *tabula rasa* and become better than the best human players (Silver et al., 2017). This was achieved by millions of games of self play where two neural networks played against each other to become iteratively better. Surprisingly, there are many similarities between Go and protein folding. One is the large number of variations that are possible, Go has 10^{172} possibilities how the black and white stones can be placed on a 19×19 board. In protein folding a similar sized complexity is given. It is estimated that for a 100 residue long protein 10^{143} conformations are possible (Levinthal, 1969), due to the large number

of degrees of freedom of the polypeptide chain. The second similarity is the move step to reach the next configuration. In Go this is comprised of moving the stone from one field to the next one on the board in each round. In protein folding the move to the next configuration would be the rotation around dihedrals ϕ and ψ. Given these similarities, the question is: could a neural network learn the folding funnel by millions of perturbations of ϕ and ψ?

In conclusion, I hope that the work presented in this thesis has contributed to advance the areas of protein-protein docking and protein folding and made it clear what limitations and problems still have to be addressed with the current methodology. Nevertheless, I am optimistic that the continued progress in computational efficiency, algorithm design and our increasing conceptual understanding of the molecular world will lead to more precise and accurate predictive methods that ultimately benefit human kind.

APPENDIX A

Supplemental Material for Chapter 2: ”Materials and Methods”

Table A.1: CASP11 and CASP12 targets overview.

Target	PDB	Description
		CASP11
TR217/T0817	4WED	Crystal structure of ABC transporter substrate-binding protein from Sinorhizobium meliloti
TR228/T0828	4Z29	Crystal structure of the magnetobacterial protein MtxA C-terminal domain
TR283/T0783	4CVH	Crystal structure of human isoprenoid synthase domain-containing protein
TR759/T0759	4Q28	Crystal Structure of the Plectin 1 and 2 Repeats of the Human Periplakin. Northeast Structural Genomics Consortium (NESG) Target HR9083A
TR760/T0760	4PQX	Crystal structure of a NigD-like protein (BACCAC_02139) from Bacteroides caccae ATCC 43185 at 2.39 A resolution
TR762/T0762	4Q5T	Crystal structure of an atmB (putative membrane lipoprotein) from Streptococcus mutans UA159 at 1.91 A resolution
TR765/T0765	4PWU	Crystal structure of a modulator protein MzrA (KPN_03524) from Klebsiella pneumoniae subsp. pneumoniae MGH 78578 at 2.45 A resolution

Table A.1: CASP

Target	PDB	Description
TR768/T0768	4OJU	Crystal structure of a leucine-rich repeat protein (BACCAP_00569) from Bacteroides capillosus ATCC 29799 at 2.00 A resolution
TR769/T0769	2MQ8	Solution NMR Structure of De novo designed protein LFR1 1 with ferredoxin fold, Northeast Structural Genomics Consortium (NESG) Target OR414
TR774/T0774	4QB7	Crystal structure of a fimbrial protein (BVU_2522) from Bacteroides vulgatus ATCC 8482 at 2.55 A resolution
TR776/T0776	4Q9A	Crystal structure of a putative GDSL-like lipase (PARMER_00689) from Parabacteroides merdae ATCC 43184 at 2.86 A resolution
TR780/T0780	4QDY	Crystal structure of a YbbR-like protein (SP_1560) from Streptococcus pneumoniae TIGR4 at 2.74 A resolution
TR782/T0782	4GRL	Crystal structure of a autoimmune TCR-MHC complex
TR783/T0783	4CVH	Crystal structure of human isoprenoid synthase domain-containing protein
TR786/T0786	4QVU	Crystal structure of a DUF4931 family protein (BCE0241) from Bacillus cereus ATCC 10987 at 2.65 A resolution
TR792/T0792	5A49	Crystal structure of the LOTUS domain (aa 139-222) of Drosophila Oskar in C222
TR795/T0795	5FJL	Crystal structure of raptor adenovirus 1 fibre head, wild-type form
TR803/T0803	4OGM	MBP-fusion protein of PilA1 residues 26-159
TR810/T0810	5JP6	Bdellovibrio bacteriovorus peptidoglycan deacetylase Bd3279
TR816/T0816	5A1Q	Crystal structure of Archaeoglobus fulgidus Af1502
TR817/T0817	4WED	Crystal structure of ABC transporter substrate-binding protein from Sinorhizobium meliloti
TR821/T0821	4R7S	Crystal structure of a tetratricopeptide repeat protein (PARMER_03812) from Parabacteroides merdae ATCC 43184 at 2.39 A resolution
TR828/T0828	4Z29	Crystal structure of the magnetobacterial protein MtxA C-terminal domain

Table A.1: CASP

Target	PDB	Description
TR829/T0829	4RQL	Crystal structure of a human cytochrome P450 2B6 (Y226H/K262R) in complex with a monoterpene - sabinene
TR833/T0833	4R03	Crystal structure of a DUF3836 family protein (BDI_3222) from Parabacteroides distasonis ATCC 8503 at 1.50 A resolution
TR837/T0837	5TF3	Crystal Structure of Protein of Unknown Function YPO2564 from Yersinia pestis
TR848/T0848	4R4Q	Crystal structure of RPA70N in complex with C31 H23 Cl2 N3 O6
TR854/T0854	4RN3	Crystal structure of a HAD-superfamily hydrolase, subfamily IA, variant 1 (GSU2069) from Geobacter sulfurreducens PCA at 2.15 A resolution
TR856/T0856	4QT6	Crystal structure of the SPRY domain of human HERC1
TR857/T0857	2MQC	NMR structure of the protein BVU_0925 from Bacteroides vulgatus ATCC 8482
		CASP12
TR862/T0862	5J5V	CdiA-CT from uropathogenic Escherichia coli in complex with cognate immunity protein and CysK
TR868/T0868	5J4A	CdiA-CT toxin from Burkholderia pseudomallei E479 in complex with cognate CdiI immunity protein
TR869/T0869	5J4A	CdiA-CT toxin from Burkholderia pseudomallei E479 in complex with cognate CdiI immunity protein
TR870/T0870	5J5V	CdiA-CT from uropathogenic Escherichia coli in complex with cognate immunity protein and CysK
TR872/T0872	5JMB	The Crystal structure of the N-terminal domain of a novel cellulases from Bacteroides coprocola
TR879/T0879	5JMU	The crystal structure of the catalytic domain of peptidoglycan N-acetylglucosamine deacetylase from Eubacterium rectale ATCC 33656
TR891/T0891	4YMP	Crystal structure of the Bacillus anthracis Hal NEAT domain in complex with heme

APPENDIX A: SUPPLEMENTAL MATERIAL FOR CHAPTER 2: "MATERIALS AND METHODS"

Table A.1: CASP

Target	PDB	Description
TR893/T0893	5IDJ	Bifunctional histidine kinase CckA (domains DHp-CA) in complex with ADP/Mg2+
TR921/T0921	5AOZ	High resolution SeMet structure of the third cohesin from Ruminococcus flavefaciens scaffoldin protein, ScaB
TR928/T0928	5TF2	CRYSTAL STRUCTURE OF THE WD40 DOMAIN OF THE HUMAN PROLACTIN REGULATORY ELEMENT-BINDING PROTEIN
TR944/T0944	5KO9	Crystal Structure of the SRAP Domain of Human HMCES Protein
TR945/T0945	5LEV	Crystal structure of human UDP-N-acetylglucosamine-dolichyl-phosphate N-acetylglucosaminephosphotransferase (DPAGT1) (V264G mutant)

Table A.2: CAPRI score_set targets overview

Target	PDB	Description
T29	2VDU	Structure of trm8-trm82, THE YEAST TRNA m7G methylation complex
T30	2REX	Crystal structure of the effector domain of PLXNB1 bound with Rnd1 GTPase
T32	3BX1	Complex between the Barley alpha-Amylase/Subtilisin Inhibitor and the subtilisin Savinase
T35	2W5F	High resolution crystallographic structure of the Clostridium thermocellum N-terminal endo-1,4-beta-D-xylanase 10B (Xyn10B) CBM22-1- GH10 modules complexed with xylohexaose
T39	3FM8	Crystal structure of full length centaurin alpha-1 bound with the FHA domain of KIF13B (CAPRI target)
T40	3E8L	The Crystal Structure of the Double-headed Arrowhead Protease Inhibitor A in Complex with Two Trypsins
T41	2WPT	The crystal structure of Im2 in complex with colicin E9 DNase
T46	3Q87	Structure of eukaryotic translation termination complex methyltransferase Mtq2-Trm112
T47	3U43	Crystal structure of the colicin E2 DNase-Im2 complex
T53	4JW2	Selection of specific protein binders for pre-defined targets from an optimized library of artificial helicoidal repeat proteins (alphaRep)
T54	4JW3	Selection of specific protein binders for pre-defined targets from an optimized library of artificial helicoidal repeat proteins (alphaRep)

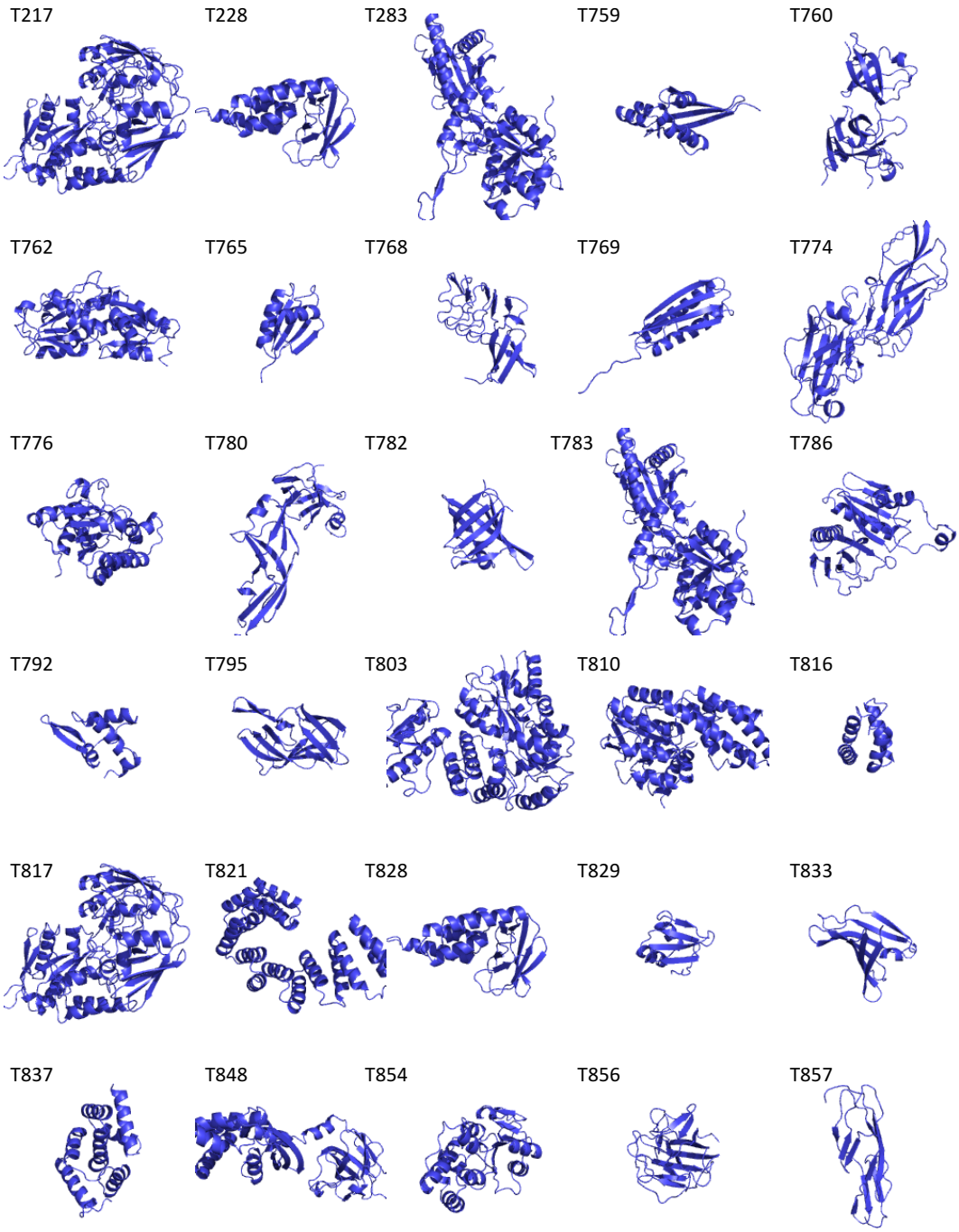

Figure A.1: CASP11

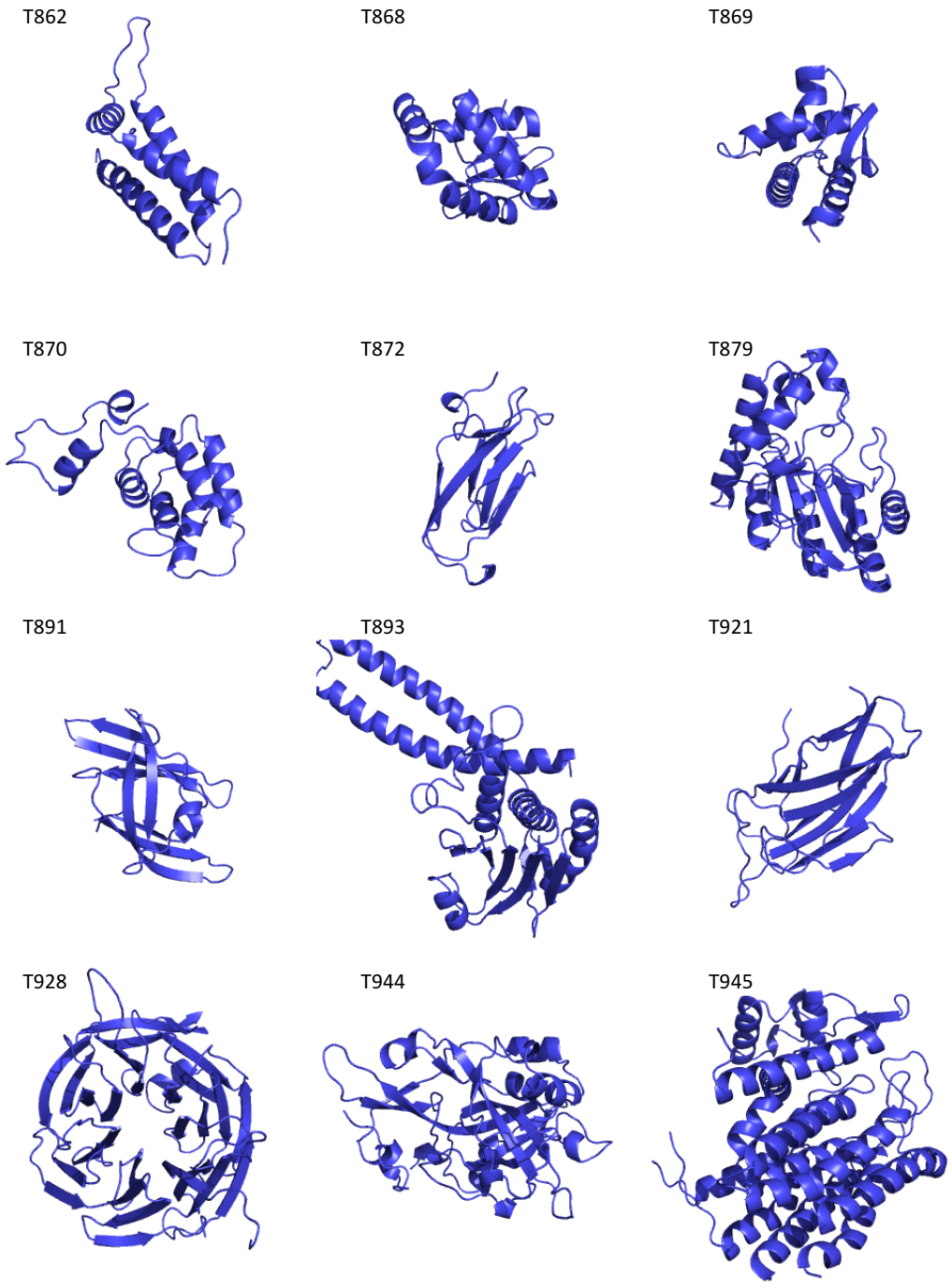

Figure A.2: CASP12

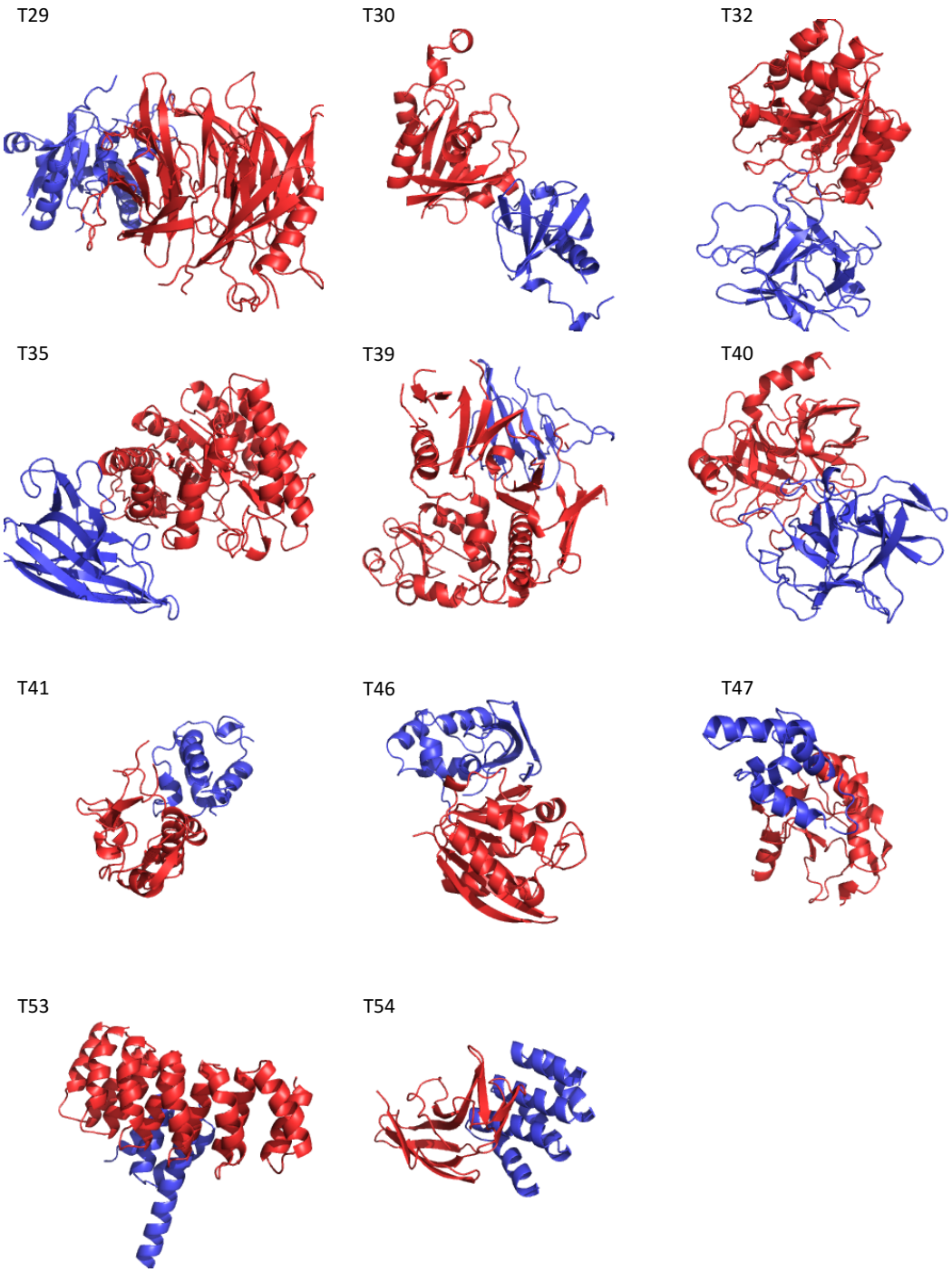

Figure A.3: CAPRI score set

APPENDIX B

Supplemental Material for Chapter 3: "A Machine Learning Approach for the Identification of Near-Native Binding Sites of Protein-Protein Complexes"

Table B.1: Protein-protein interaction molecular descriptor list and features. All used features are shown with its descriptor name, the descriptor category (desc. cat, see Table 2.1 for description of these), the feature category (feat. cat.) and a reference. The feature categories are cd (cluster distribution) and cc (cluster count). The cd has the features associated that describe the minimum (MIN), maximum (MAX), median (AVG), first quartile (Q1) and third quartile (Q3) of the cluster distribution. The cc category has two features C1 and C2, for number of models in cluster 1 and 2 in a pair-wise comparison, respectively.

Descriptor	**Desc. Cat.**	**Feat. Cat.**	**Reference**
N_AP_DOPE_HR	ac	cd	Shen and Sali (2006)
N_AP_W1	ac	cd	Mintseris et al. (2007)
N_AP_DOPE	ac	cd	Shen and Sali (2006)
N_AP_OPUS_PSP	ac	cd	Xu et al. (2017)
N_AP_DCOMPLEX	ac	cd	Liu et al. (2004)
N_AP_GOAP_ALL	ac	cd	Zhou and Skolnick (2011)

Table B.1: Protein-protein interaction molecular descriptor list and features. All used features are shown with its descriptor name, the descriptor category (desc. cat, see Table 2.1 for description of these), the feature category (feat. cat.) and a reference. The feature categories are cd (cluster distribution) and cc (cluster count). The cd has the features associated that describe the minimum (MIN), maximum (MAX), median (AVG), first quartile (Q1) and third quartile (Q3) of the cluster distribution. The cc category has two features C1 and C2, for number of models in cluster 1 and 2 in a pair-wise comparison, respectively.

Descriptor	**Desc. Cat.**	**Feat. Cat.**	**Reference**
N_AP_GOAP_G	ac	cd	Zhou and Skolnick (2011)
N_AP_dDFIRE	ac	cd	Yang and Zhou (2008)
N_AP_MPS	ac	cd	Chuang et al. (2008)
N_AP_calRW	ac	cd	Zhang and Zhang (2010)
N_AP_T2	ac	cd	Tobi (2010)
N_AP_T1	ac	cd	Tobi (2010)
N_AP_ACE	ac	cd	Andrusier et al. (2007)
N_DCOMPLEX	ac	cd	Liu et al. (2004)
N_AP_URS	ac	cd	Chuang et al. (2008)
N_AP_DFIRE2	ac	cd	Yang and Zhou (2008)
N_AP_GOAP_DF	ac	cd	Zhou and Skolnick (2011)
N_AP_DARS	ac	cd	Chuang et al. (2008)
N_AP_calRWp	ac	cd	Zhang and Zhang (2010)
N_AP_DDG_W	ac	cd	Nguyen et al. (2013)
N_AP_DDG_U	ac	cd	Nguyen et al. (2013)
N_FIREDOCK_EI	cs	cd	Andrusier et al. (2007)
N_ZRANK2	cs	cd	Pierce and Weng (2008)
N_PYDOCK_TOT	cs	cd	Cheng et al. (2007)
N_ZRANK_y	cs	cd	Pierce and Weng (2007)
N_ROSETTADOCK	cs	cd	Chaudhury et al. (2010)
N_ZRANK_x	cs	cd	Pierce and Weng (2007)
N_SIPPER	cs	cd	Pons et al. (2011)
N_FIREDOCK_AB	cs	cd	Andrusier et al. (2007)
N_CP_PIE	cs	cd	Ravikant and Elber (2010)
N_AP_PISA	cs	cd	Viswanath et al. (2013)
N_FIREDOCK	cs	cd	Andrusier et al. (2007)
N_HBOND	hb	cd	Andrusier et al. (2007)
N_HBOND2	hb	cd	Chaudhury et al. (2010)
N_NHB	hb	cd	Chaudhury et al. (2010)
N_AA_PROP	mi	cd	Chaudhury et al. (2010)
N_NIPacking	mi	cd	Mitra and Pal (2010)
N_ALIPH	mi	cd	Andrusier et al. (2007)
N_AP_GEOMETRIC	mi	cd	not published
N_DDG_V	mi	cd	not published

Table B.1: Protein-protein interaction molecular descriptor list and features. All used features are shown with its descriptor name, the descriptor category (desc. cat, see Table 2.1 for description of these), the feature category (feat. cat.) and a reference. The feature categories are cd (cluster distribution) and cc (cluster count). The cd has the features associated that describe the minimum (MIN), maximum (MAX), median (AVG), first quartile (Q1) and third quartile (Q3) of the cluster distribution. The cc category has two features C1 and C2, for number of models in cluster 1 and 2 in a pair-wise comparison, respectively.

Descriptor	**Desc. Cat.**	**Feat. Cat.**	**Reference**
N_TRANS_S	mi	cd	not published
N_PI_PI	mi	cd	Andrusier et al. (2007)
N_INSIDE	mi	cd	Andrusier et al. (2007)
N_ROT_S	mi	cd	not published
N_NSC	mi	cd	Mitra and Pal (2010)
N_CAT_PI	mi	cd	Andrusier et al. (2007)
N_CP_SKOIP	rc	cd	Lu et al. (2003)
N_CP_TSC	rc	cd	Tobi (2010)
N_CP_VD	rc	cd	Pokarowski et al. (2005); Feng et al. (2010)
N_CP_Qp	rc	cd	Pokarowski et al. (2005); Feng et al. (2010)
N_CP_DDG_U	rc	cd	Nguyen et al. (2013)
N_CP_DDG_W	rc	cd	Nguyen et al. (2013)
N_CP_Qa	rc	cd	Pokarowski et al. (2005); Feng et al. (2010)
N_CP_MJPL	rc	cd	Pokarowski et al. (2005); Feng et al. (2010)
N_CP_BFKV	rc	cd	Pokarowski et al. (2005); Feng et al. (2010)
N_CP_SKOa	rc	cd	Pokarowski et al. (2005); Feng et al. (2010)
N_CP_SKOb	rc	cd	Pokarowski et al. (2005); Feng et al. (2010)
N_CP_MJ2h	rc	cd	Pokarowski et al. (2005); Feng et al. (2010)
N_CP_SJKG	rc	cd	Pokarowski et al. (2005); Feng et al. (2010)
N_CP_MJ2	rc	cd	Pokarowski et al. (2005); Feng et al. (2010)
N_CP_MJ1	rc	cd	Pokarowski et al. (2005); Feng et al. (2010)
N_CP_D1	rc	cd	Liu and Vakser (2011)
N_CP_GKS	rc	cd	Pokarowski et al. (2005); Feng et al. (2010)

Table B.1: Protein-protein interaction molecular descriptor list and features. All used features are shown with its descriptor name, the descriptor category (desc. cat, see Table 2.1 for description of these), the feature category (feat. cat.) and a reference. The feature categories are cd (cluster distribution) and cc (cluster count). The cd has the features associated that describe the minimum (MIN), maximum (MAX), median (AVG), first quartile (Q1) and third quartile (Q3) of the cluster distribution. The cc category has two features C1 and C2, for number of models in cluster 1 and 2 in a pair-wise comparison, respectively.

Descriptor	Desc. Cat.	Feat. Cat.	Reference
N_CP_HLPL	rc	cd	Pokarowski et al. (2005); Feng et al. (2010)
N_CP_BT	rc	cd	Pokarowski et al. (2005); Feng et al. (2010)
N_CP_RMFCEN1	rc	cd	Rajgaria et al. (2008)
N_CP_RMFCEN2	rc	cd	Rajgaria et al. (2008)
N_CP_MJ3h	rc	cd	Pokarowski et al. (2005); Feng et al. (2010)
N_CP_BL	rc	cd	Pokarowski et al. (2005); Feng et al. (2010)
N_CP_MSBM	rc	cd	Pokarowski et al. (2005); Feng et al. (2010)
N_PROPNSTS	rc	cd	Pons et al. (2011)
N_CP_TEl	rc	cd	Pokarowski et al. (2005); Feng et al. (2010)
N_CP_RO	rc	cd	Pokarowski et al. (2005); Feng et al. (2010)
N_CP_TEs	rc	cd	Pokarowski et al. (2005); Feng et al. (2010)
N_CP_TB	rc	cd	Tobi and Bahar (2006)
N_CP_TD	rc	cd	Pokarowski et al. (2005); Feng et al. (2010)
N_CP_MS	rc	cd	Pokarowski et al. (2005); Feng et al. (2010)
N_CP_TS	rc	cd	Pokarowski et al. (2005); Feng et al. (2010)
N_CP_RMFCA	rc	cd	Rajgaria et al. (2006)
N_CP_Qm	rc	cd	Pokarowski et al. (2005); Feng et al. (2010)
N_CG_ENV	se	cd	Chaudhury et al. (2010)
N_LK_SOLV	se	cd	Chaudhury et al. (2010)
N_CG_BETA	se	cd	Chaudhury et al. (2010)
N_DESOLV	se	cd	Cheng et al. (2007)
N_ODA	se	cd	FernandezRecio et al. (2005)

Table B.1: Protein-protein interaction molecular descriptor list and features. All used features are shown with its descriptor name, the descriptor category (desc. cat, see Table 2.1 for description of these), the feature category (feat. cat.) and a reference. The feature categories are cd (cluster distribution) and cc (cluster count). The cd has the features associated that describe the minimum (MIN), maximum (MAX), median (AVG), first quartile (Q1) and third quartile (Q3) of the cluster distribution. The cc category has two features C1 and C2, for number of models in cluster 1 and 2 in a pair-wise comparison, respectively.

Descriptor	**Desc. Cat.**	**Feat. Cat.**	**Reference**
N_CP_ZLOCAL_MIN	sp	cd	Feliu et al. (2011)
N_CP_ELOCAL_MIN	sp	cd	Feliu et al. (2011)
N_CP_E3DC_CB	sp	cd	Feliu et al. (2011)
N_CP_ZPAIR_CB	sp	cd	Feliu et al. (2011)
N_CP_Z3DC_CB	sp	cd	Feliu et al. (2011)
N_CP_Z3DC_MIN	sp	cd	Feliu et al. (2011)
N_CP_ZLOCAL_CB	sp	cd	Feliu et al. (2011)
N_CP_ES3DC_MIN	sp	cd	Feliu et al. (2011)
N_CP_ELOCAL_CB	sp	cd	Feliu et al. (2011)
N_CP_ZS3DC_MIN	sp	cd	Feliu et al. (2011)
N_CP_EPAIR_CB	sp	cd	Feliu et al. (2011)
N_CP_ZS3DC_CB	sp	cd	Feliu et al. (2011)
N_CP_E3D_CB	sp	cd	Feliu et al. (2011)
N_CP_ZPAIR_MIN	sp	cd	Feliu et al. (2011)
N_CP_E3D_MIN	sp	cd	Feliu et al. (2011)
N_CP_ES3DC_CB	sp	cd	Feliu et al. (2011)
N_CP_EPAIR_MIN	sp	cd	Feliu et al. (2011)
N_CP_E3DC_MIN	sp	cd	Feliu et al. (2011)
N_ELE	ve	cd	Cheng et al. (2007)
N_VDW	ve	cd	Cheng et al. (2007)
N_FA_REP	ve	cd	Chaudhury et al. (2010)
N_CG_VDW	ve	cd	Chaudhury et al. (2010)
N_CG_PP	ve	cd	Feliu et al. (2011)
N_FA_PP	ve	cd	Chaudhury et al. (2010)
COUNT	NA	cc	NA

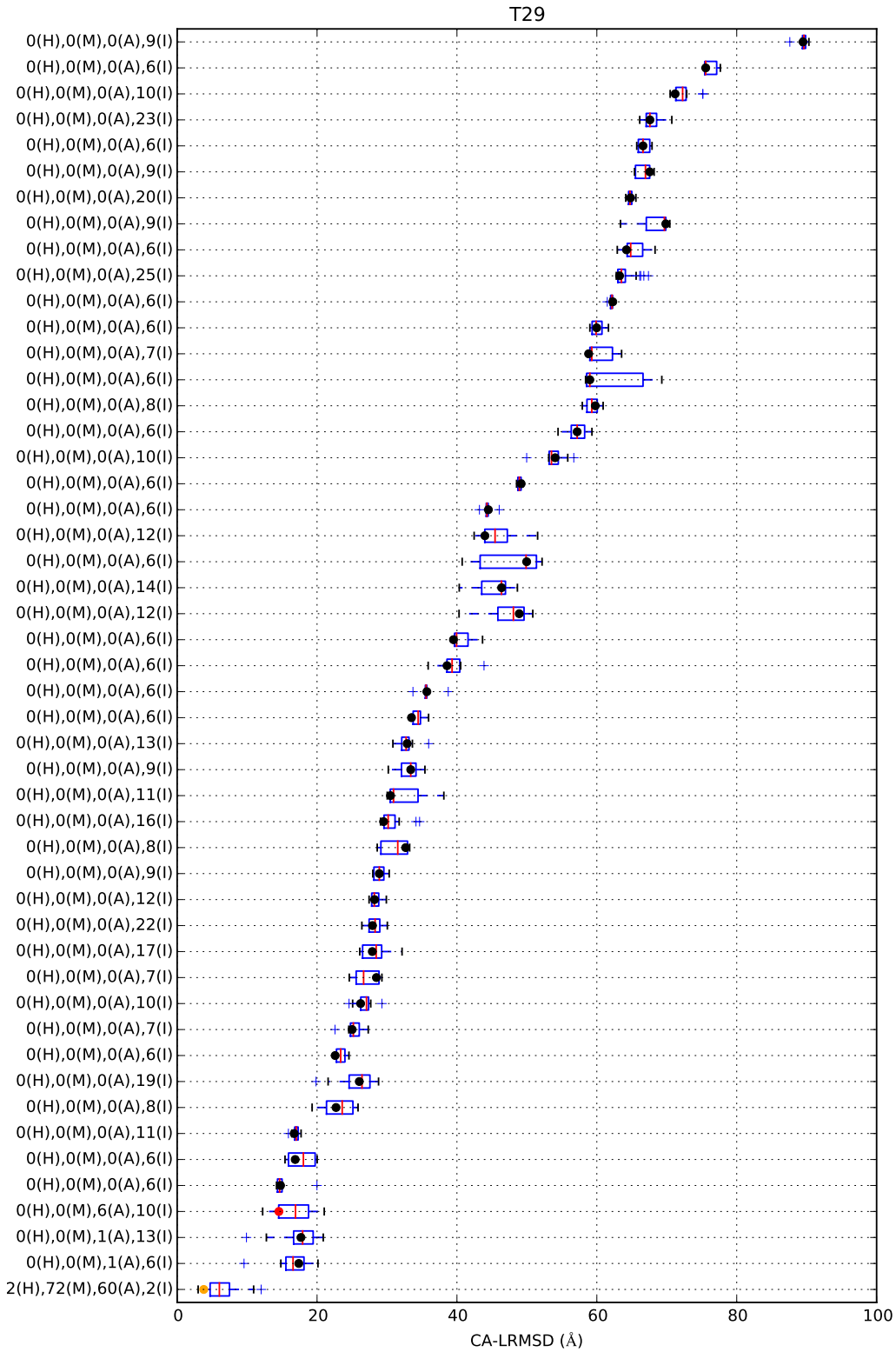

Figure B.1: T29; Cα-LRMSD distribution of all clusters. Labels on the y-axis indicate the number of high (H), medium (M), acceptable (A) and incorrect (I) solutions per cluster. The coloured sphere (green: high; orange: medium; red: acceptable; black: incorrect) in each boxplot indicates the Cα-LRMSD of the cluster centroid. Permission to reproduce this Figure has been granted by John Wiley & Sons, Inc.

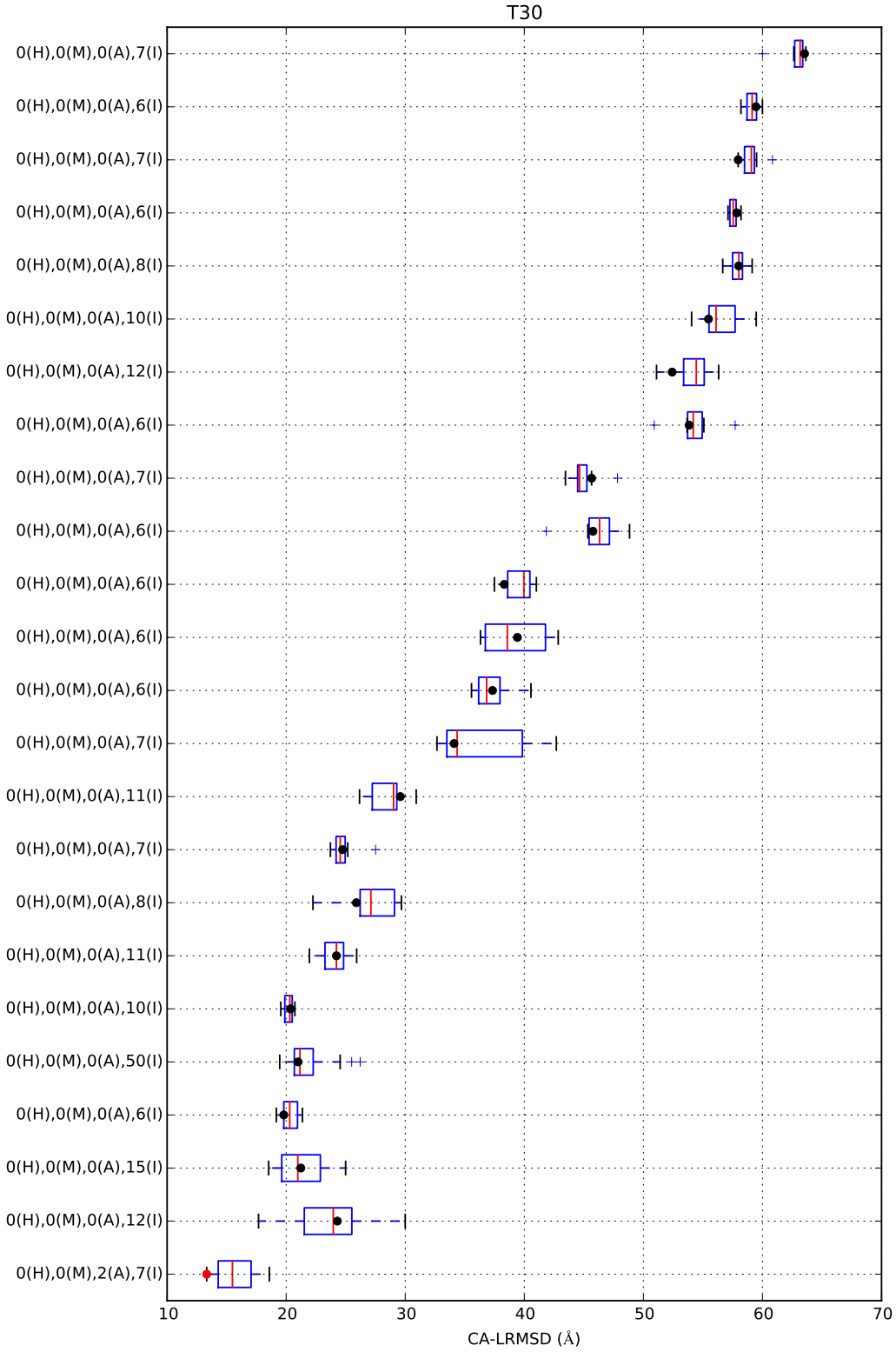

Figure B.2: T30; Cα-LRMSD distribution of all clusters. Labels on the y-axis indicate the number of high (H), medium (M), acceptable (A) and incorrect (I) solutions per cluster. The colored sphere (green: high; orange: medium; red: acceptable; black: incorrect) in each boxplot indicates the Cα-LRMSD of the cluster centroid. Permission to reproduce this Figure has been granted by John Wiley & Sons, Inc.

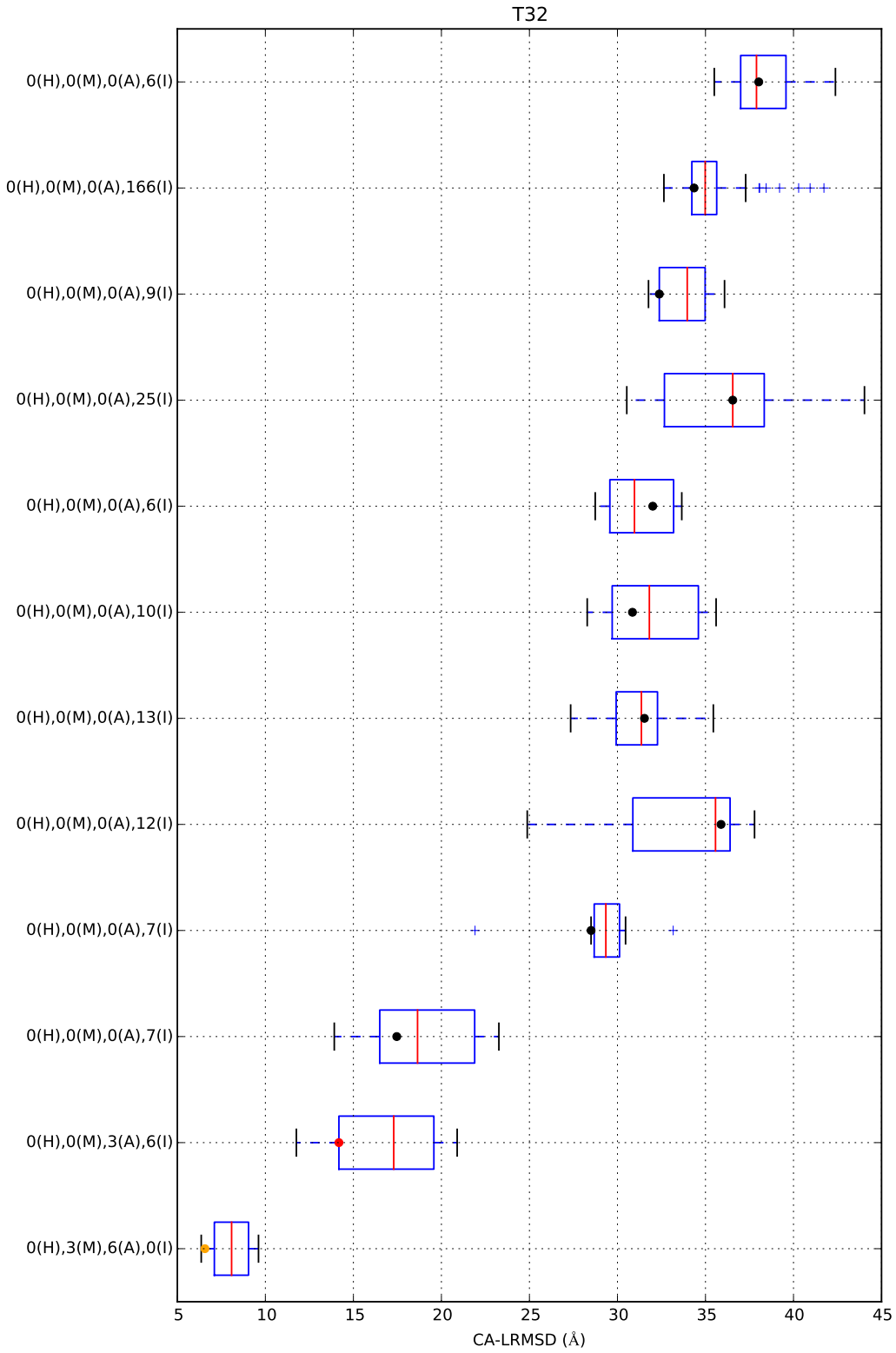

Figure B.3: T32; Cα-LRMSD distribution of all clusters. Labels on the y-axis indicate the number of high (H), medium (M), acceptable (A) and incorrect (I) solutions per cluster. The colored sphere (green: high; orange: medium; red: acceptable; black: incorrect) in each boxplot indicates the Cα-LRMSD of the cluster centroid. Permission to reproduce this Figure has been granted by John Wiley & Sons, Inc.

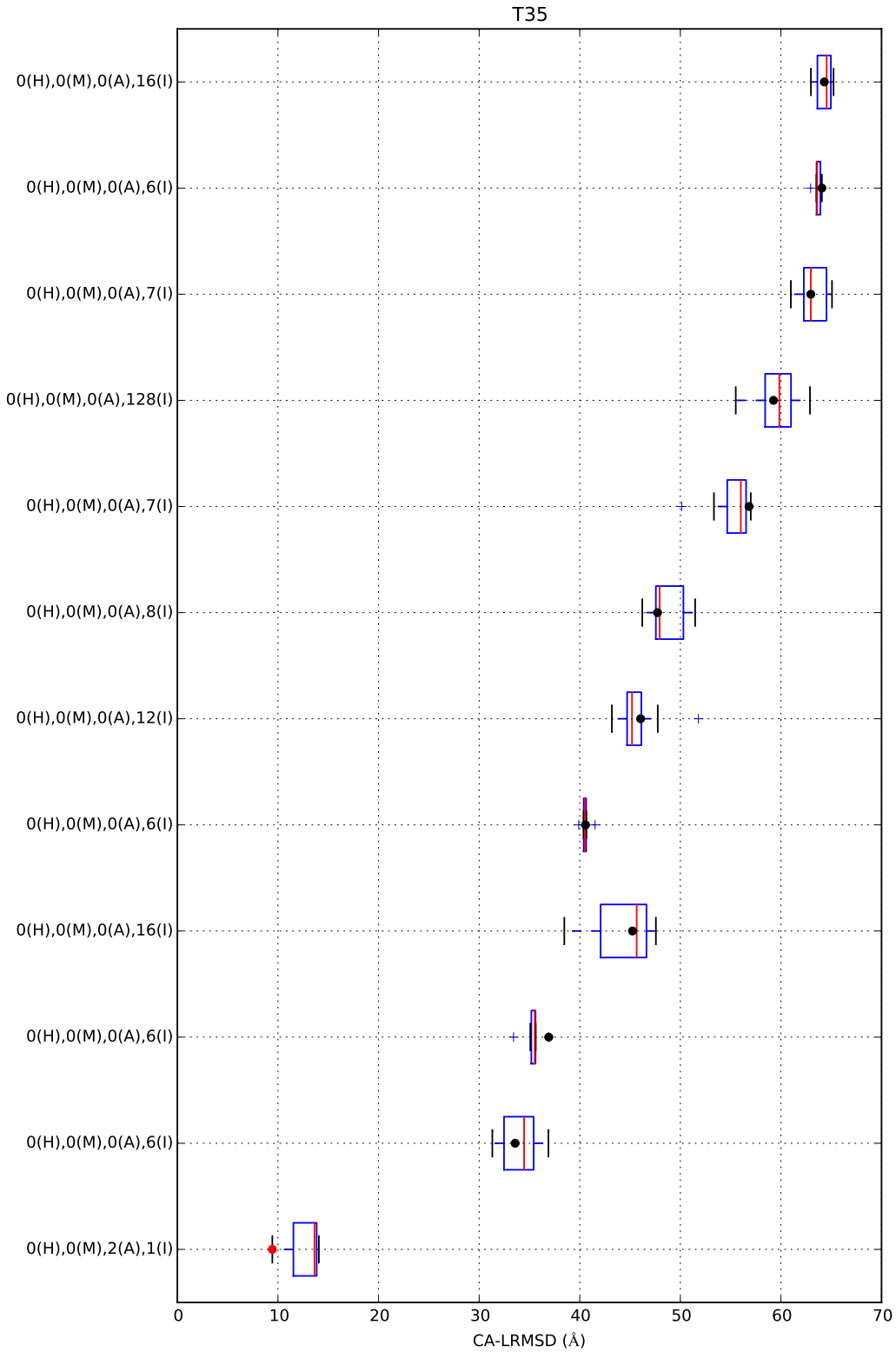

Figure B.4: T35; Cα-LRMSD distribution of all clusters. Labels on the y-axis indicate the number of high (H), medium (M), acceptable (A) and incorrect (I) solutions per cluster. The colored sphere (green: high; orange: medium; red: acceptable; black: incorrect) in each boxplot indicates the Cα-LRMSD of the cluster centroid. Permission to reproduce this Figure has been granted by John Wiley & Sons, Inc.

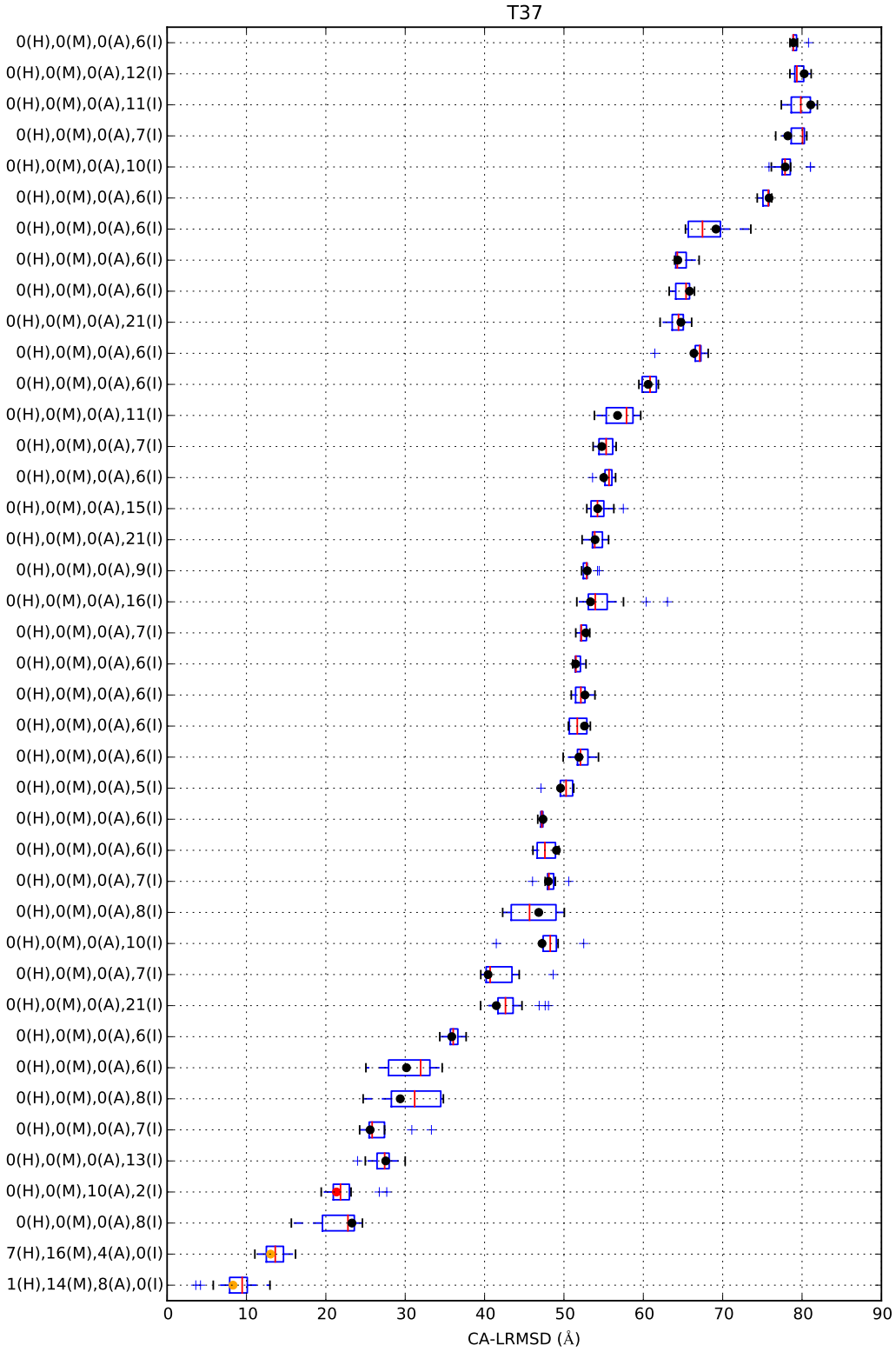

Figure B.5: T37; Cα-LRMSD distribution of all clusters. Labels on the y-axis indicate the number of high (H), medium (M), acceptable (A) and incorrect (I) solutions per cluster. The colored sphere (green: high; orange: medium; red: acceptable; black: incorrect) in each boxplot indicates the Cα-LRMSD of the cluster centroid. Permission to reproduce this Figure has been granted by John Wiley & Sons, Inc.

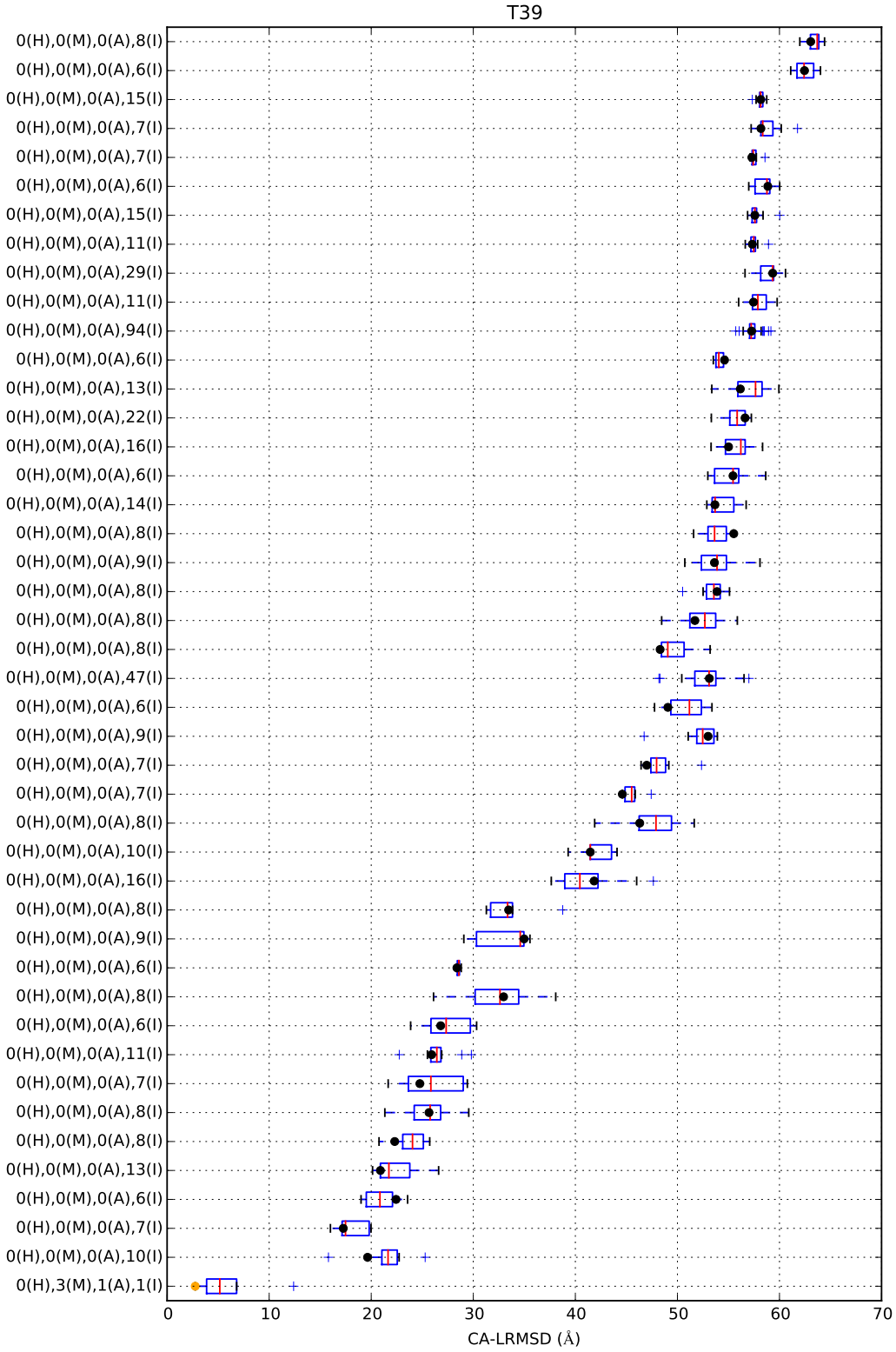

Figure B.6: T39; Cα-LRMSD distribution of all clusters. Labels on the y-axis indicate the number of high (H), medium (M), acceptable (A) and incorrect (I) solutions per cluster. The colored sphere (green: high; orange: medium; red: acceptable; black: incorrect) in each boxplot indicates the Cα-LRMSD of the cluster centroid. Permission to reproduce this Figure has been granted by John Wiley & Sons, Inc.

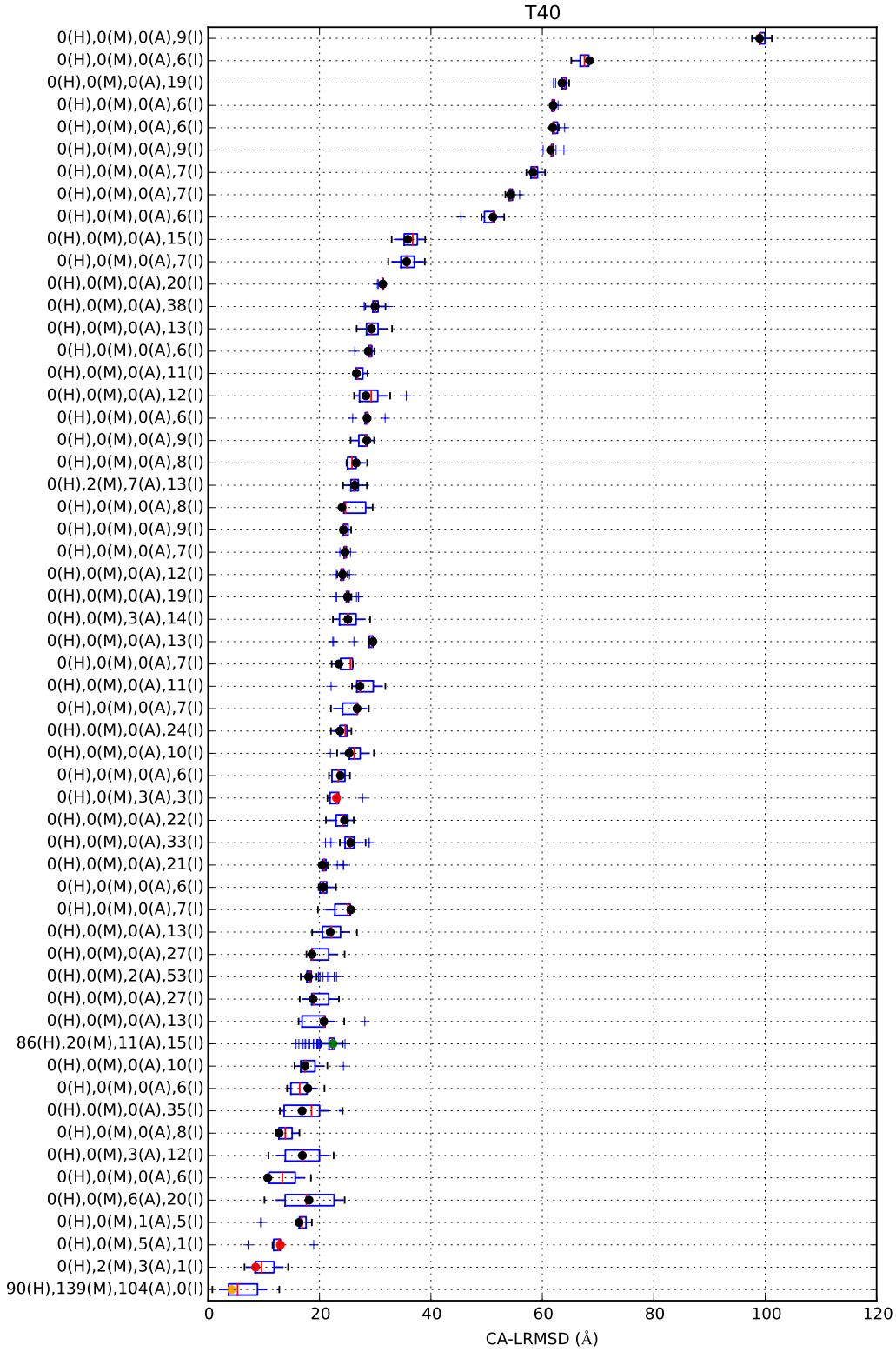

Figure B.7: T40; Cα-LRMSD distribution of all clusters. Labels on the y-axis indicate the number of high (H), medium (M), acceptable (A) and incorrect (I) solutions per cluster. The colored sphere (green: high; orange: medium; red: acceptable; black: incorrect) in each boxplot indicates the Cα-LRMSD of the cluster centroid. Permission to reproduce this Figure has been granted by John Wiley & Sons, Inc.

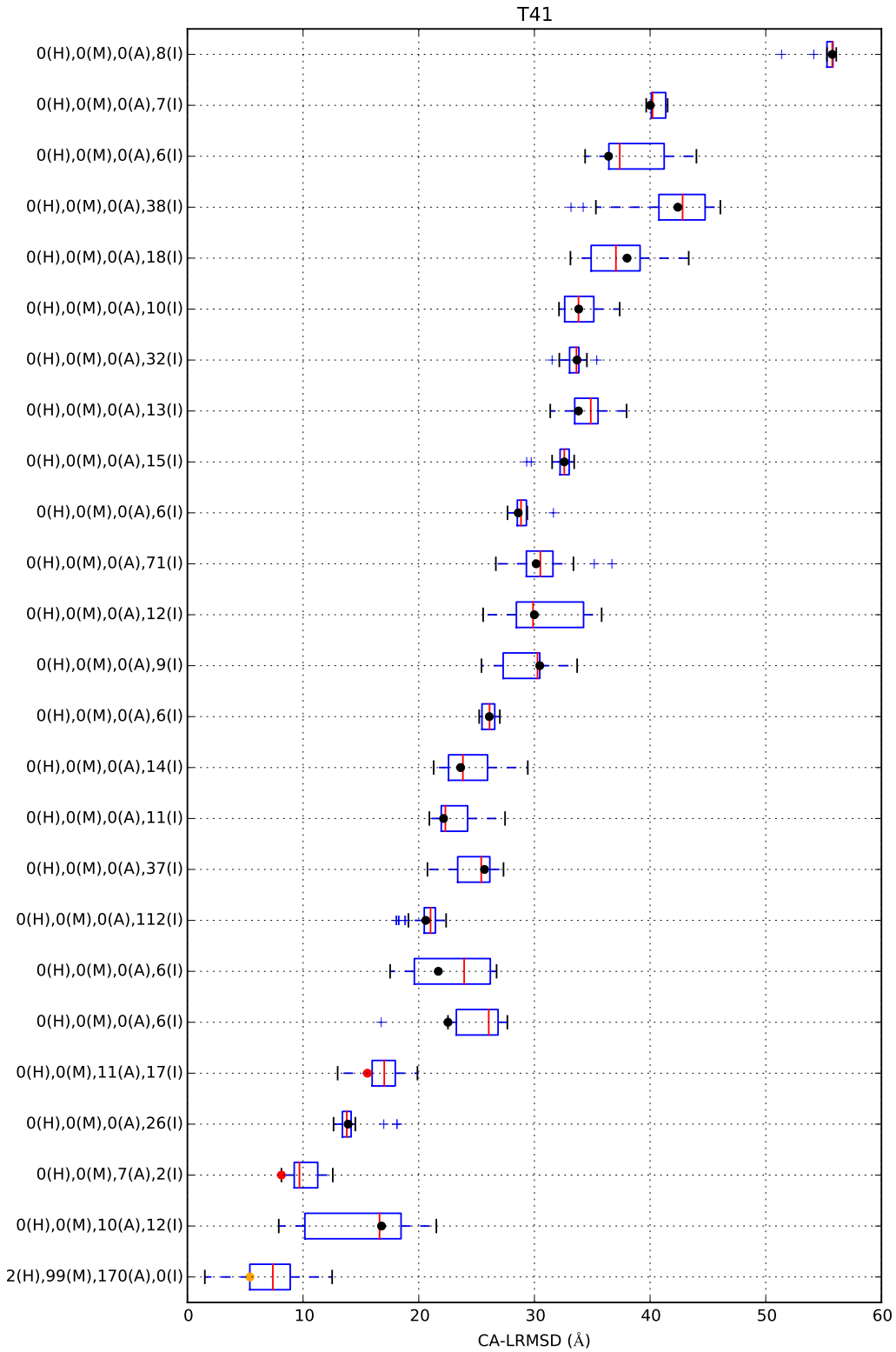

Figure B.8: T41; Cα-LRMSD distribution of all clusters. Labels on the y-axis indicate the number of high (H), medium (M), acceptable (A) and incorrect (I) solutions per cluster. The colored sphere (green: high; orange: medium; red: acceptable; black: incorrect) in each boxplot indicates the Cα-LRMSD of the cluster centroid. Permission to reproduce this Figure has been granted by John Wiley & Sons, Inc.

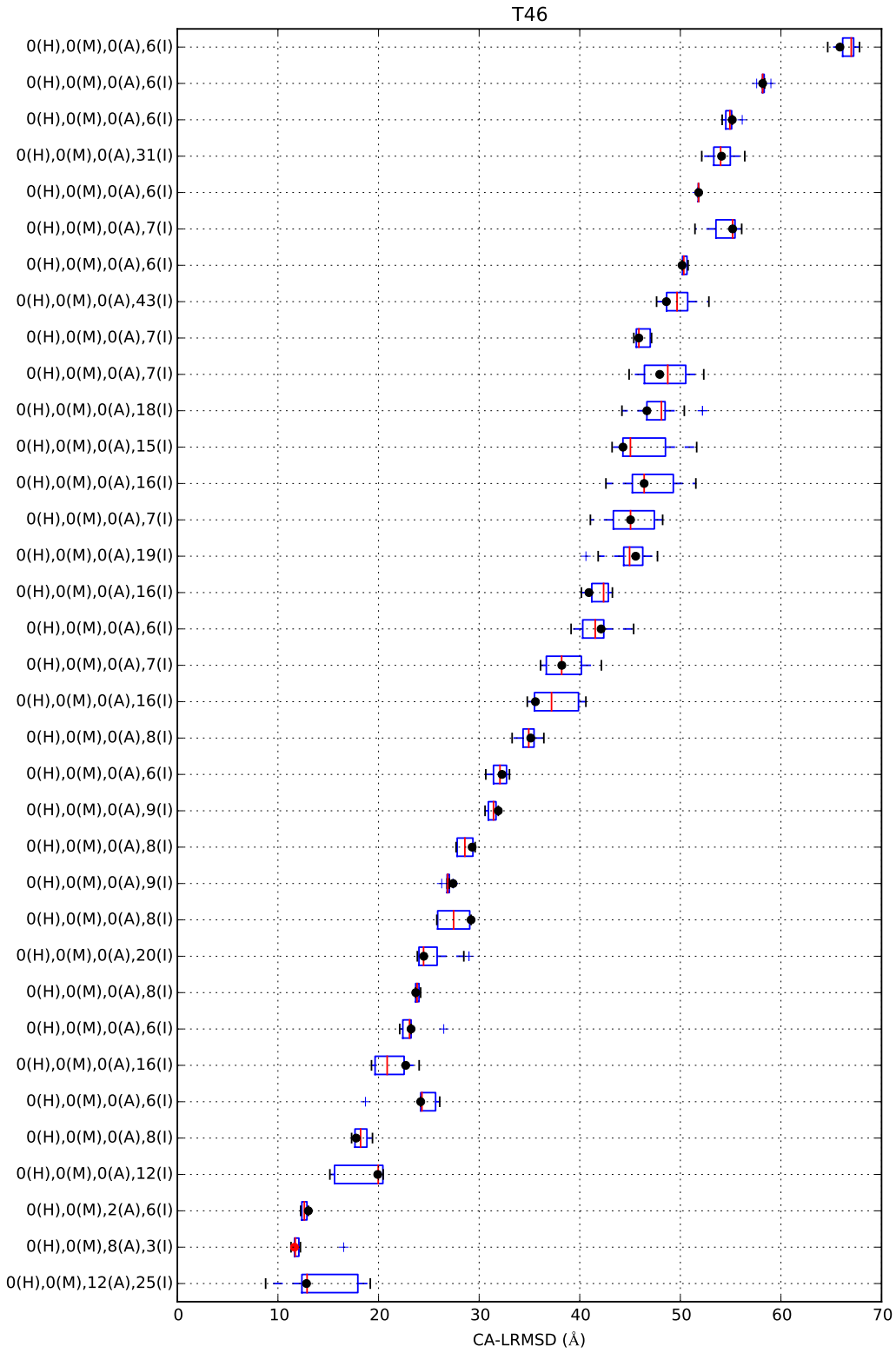

Figure B.9: T46; Cα-LRMSD distribution of all clusters. Labels on the y-axis indicate the number of high (H), medium (M), acceptable (A) and incorrect (I) solutions per cluster. The colored sphere (green: high; orange: medium; red: acceptable; black: incorrect) in each boxplot indicates the Cα-LRMSD of the cluster centroid. Permission to reproduce this Figure has been granted by John Wiley & Sons, Inc.

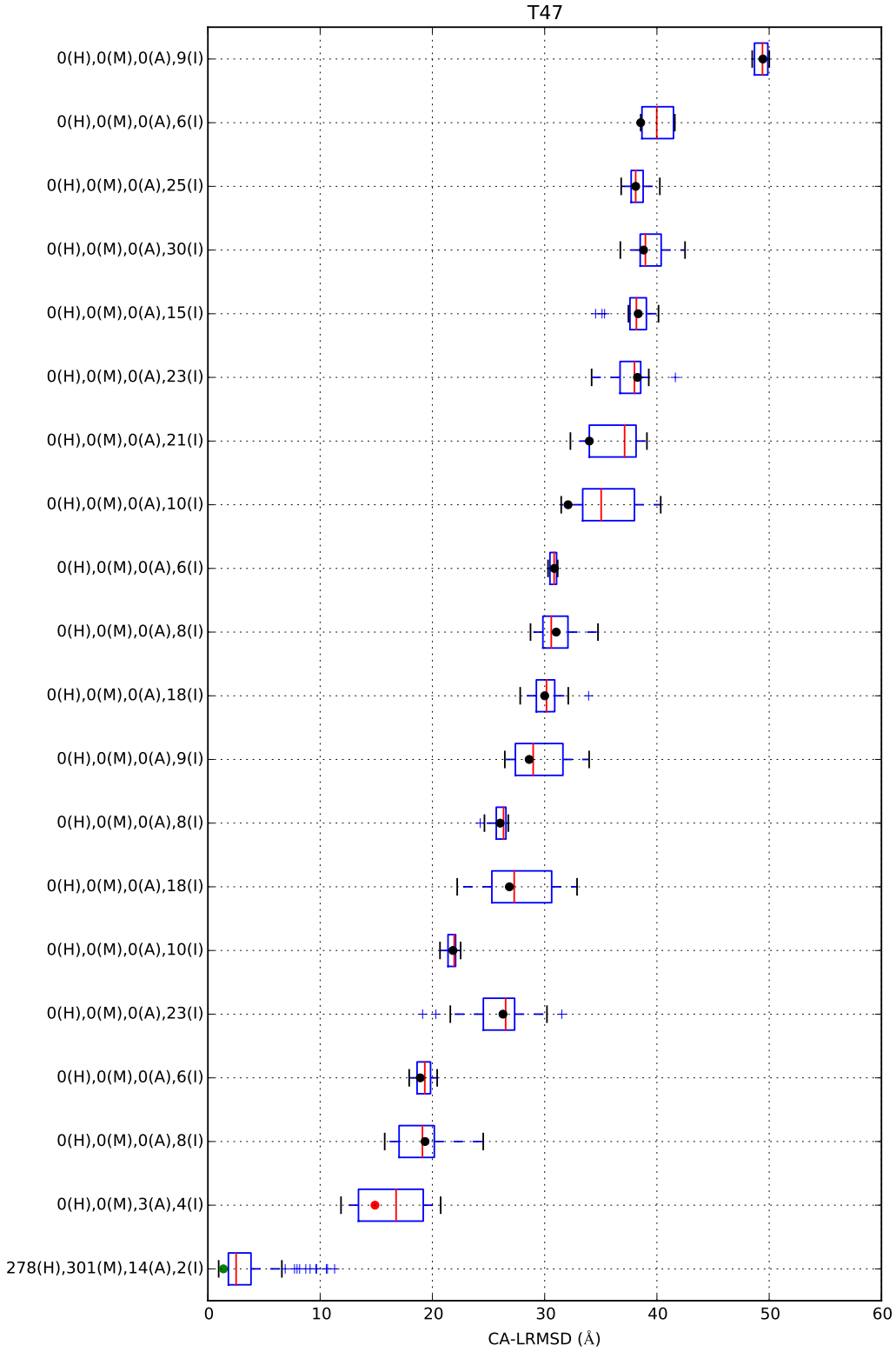

Figure B.10: T47; Cα-LRMSD distribution of all clusters. Labels on the y-axis indicate the number of high (H), medium (M), acceptable (A) and incorrect (I) solutions per cluster. The colored sphere (green: high; orange: medium; red: acceptable; black: incorrect) in each boxplot indicates the Cα-LRMSD of the cluster centroid. Permission to reproduce this Figure has been granted by John Wiley & Sons, Inc.

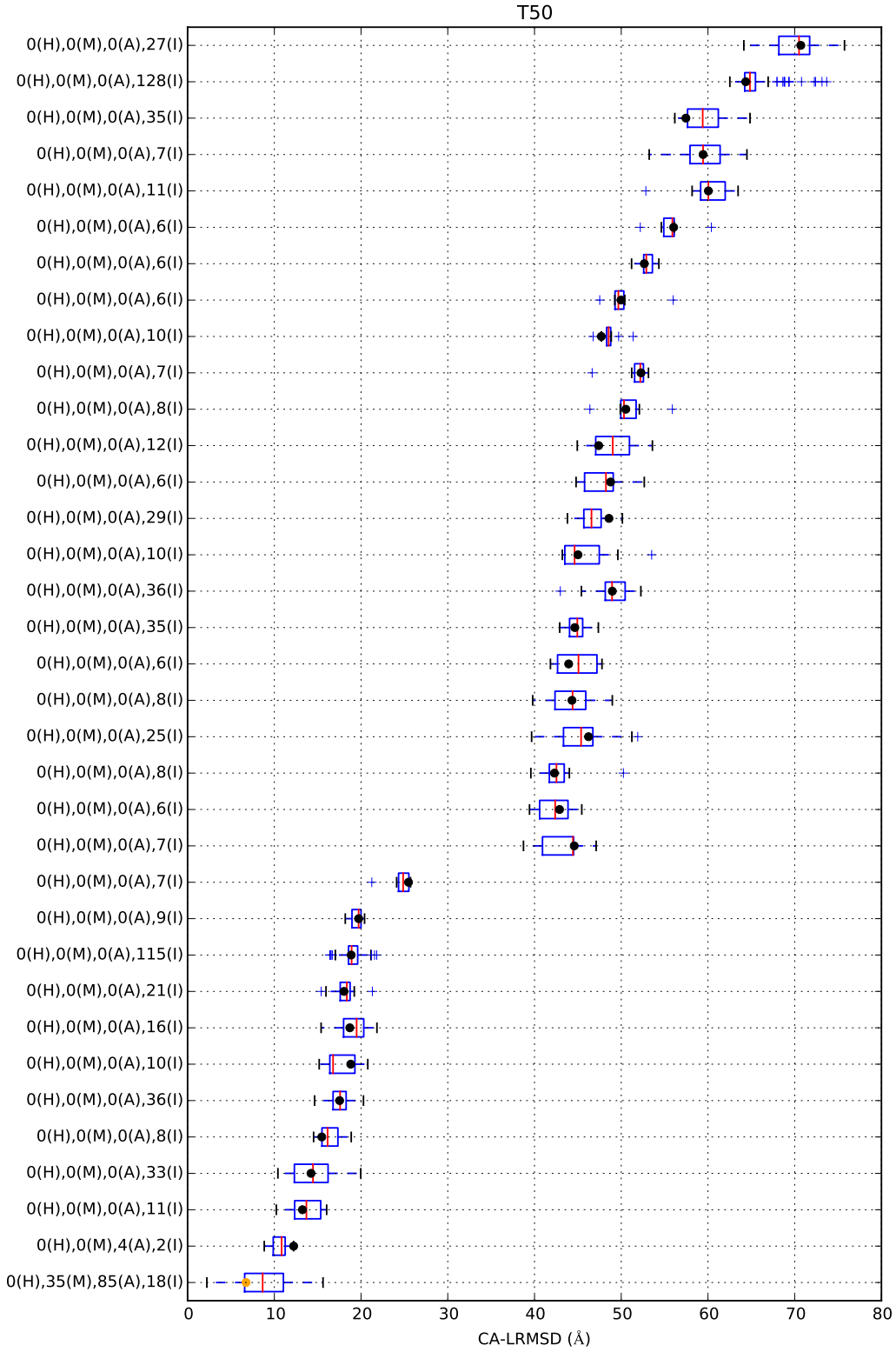

Figure B.11: T50; Cα-LRMSD distribution of all clusters. Labels on the y-axis indicate the number of high (H), medium (M), acceptable (A) and incorrect (I) solutions per cluster. The colored sphere (green: high; orange: medium; red: acceptable; black: incorrect) in each boxplot indicates the Cα-LRMSD of the cluster centroid. Permission to reproduce this Figure has been granted by John Wiley & Sons, Inc.

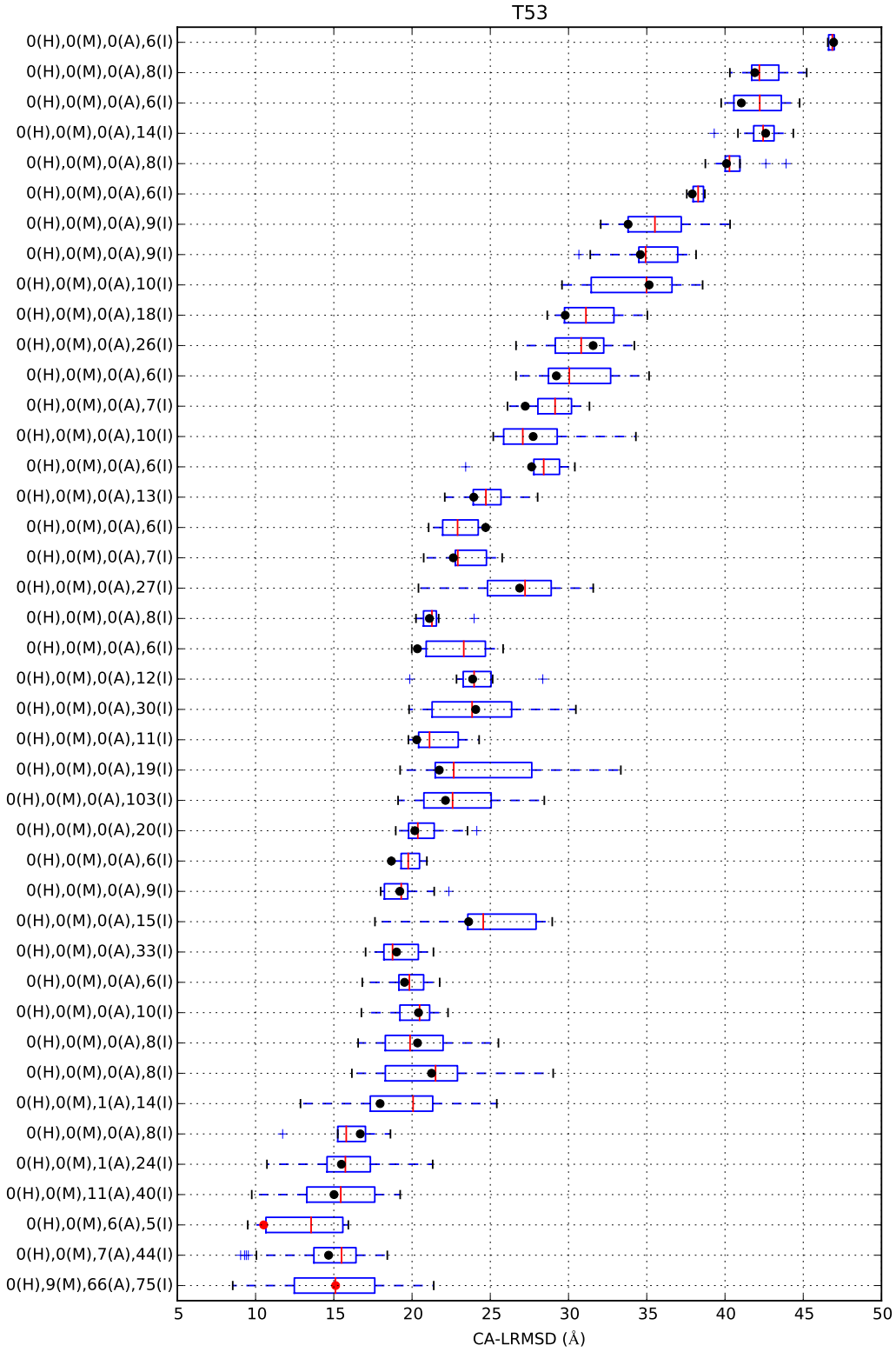

Figure B.12: T53; Cα-LRMSD distribution of all clusters. Labels on the y-axis indicate the number of high (H), medium (M), acceptable (A) and incorrect (I) solutions per cluster. The colored sphere (green: high; orange: medium; red: acceptable; black: incorrect) in each boxplot indicates the Cα-LRMSD of the cluster centroid. Permission to reproduce this Figure has been granted by John Wiley & Sons, Inc.

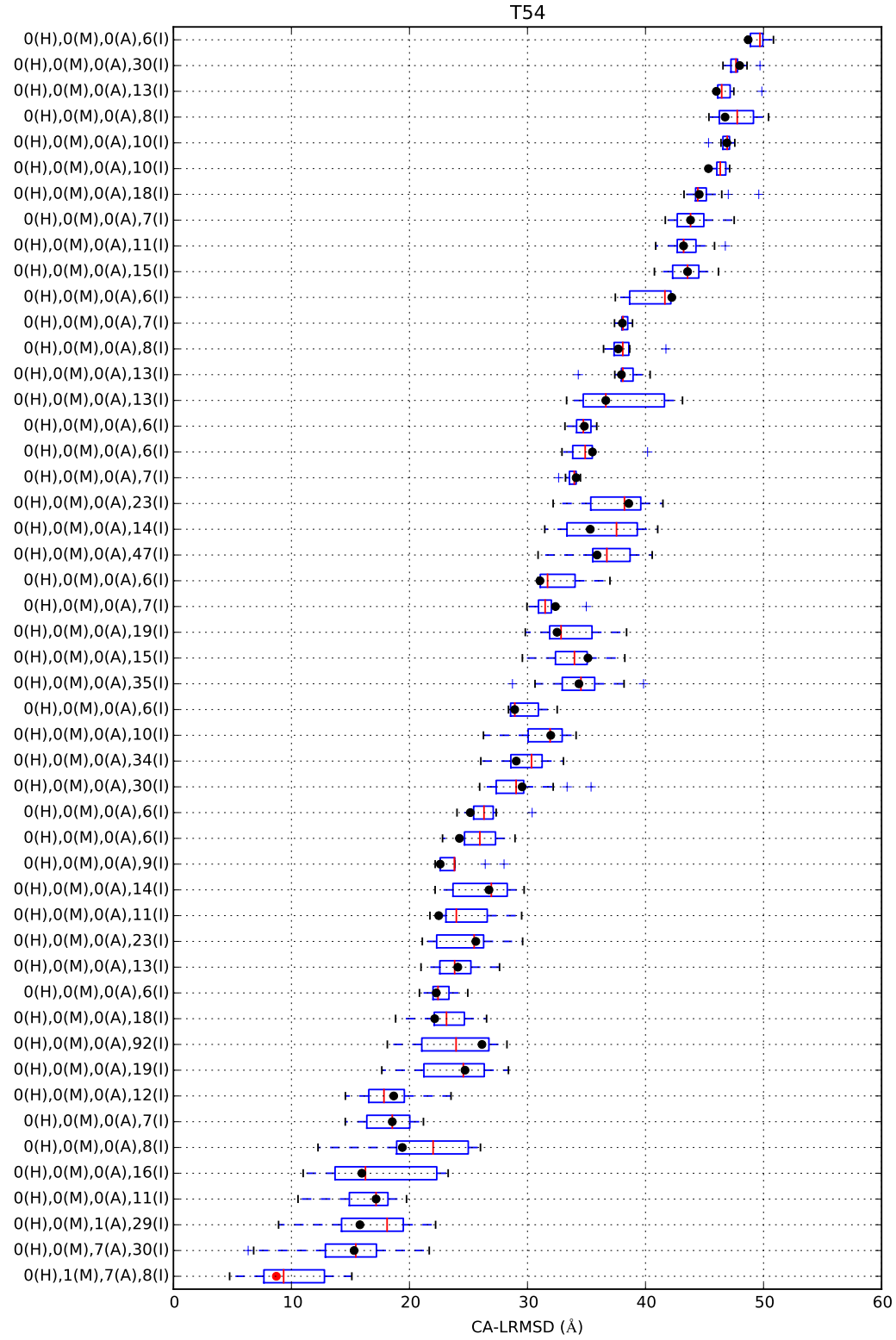

Figure B.13: T54; Cα-LRMSD distribution of all clusters. Labels on the y-axis indicate the number of high (H), medium (M), acceptable (A) and incorrect (I) solutions per cluster. The colored sphere (green: high; orange: medium; red: acceptable; black: incorrect) in each boxplot indicates the Cα-LRMSD of the cluster centroid. Permission to reproduce this Figure has been granted by John Wiley & Sons, Inc.

APPENDIX C

Supplemental Material for Chapter 4: ”Optimization of Predicted Protein Folds by Refinement”

Table C.1: Best snapshot rank. The table shows the rank of the best snapshot with respect to GDTHA for each target from the CASP11 and CASP12 benchmark set. The rank was determined by DDFIRE score. The columns SM and Best shows the starting model and Best snapshot GDTHA. Column Total shows the total number of snapshots for a target.

Target	SM	Best	Source	Rank	Total
TR217	0.628	0.656	PR	9152	34424
TR228	0.548	0.664	NR	5008	34424
TR283	0.412	0.476	DR	26523	34424
TR759	0.430	0.648	PR	5753	34424
TR760	0.573	0.608	PR	2158	34424
TR762	0.706	0.698	PR	6323	34424
TR765	0.579	0.786	NR	333	34424
TR768	0.640	0.701	PR	7564	34424
TR769	0.562	0.680	NR	18839	23216
TR774	0.381	0.401	PR	24373	34424
TR776	0.631	0.666	PR	8578	34424
TR780	0.540	0.668	DR	9874	34424
TR782	0.648	0.686	PR	6592	22416
TR783	0.586	0.638	PR	5612	34424
TR786	0.479	0.527	PR	5842	34424
TR792	0.607	0.770	DR	9669	34424
TR795	0.586	0.645	DR	9353	23216

APPENDIX C: SUPPLEMENTAL MATERIAL FOR CHAPTER 4: "OPTIMIZATION OF PREDICTED PROTEIN FOLDS BY REFINEMENT"

Table C.1: Best snapshot rank. The table shows the rank of the best snapshot with respect to GDTHA for each target from the CASP11 and CASP12 benchmark set. The rank was determined by DDFIRE score. The columns SM and Best shows the starting model and Best snapshot GDTHA. Column Total shows the total number of snapshots for a target.

Target	SM	Best	Source	Rank	Total
TR803	0.330	0.394	DR	5032	34424
TR810	0.540	0.570	PR	16136	34424
TR816	0.515	0.651	DR	4378	34424
TR817	0.468	0.564	DR	26159	34424
TR821	0.483	0.721	NR	3030	23216
TR828	0.491	0.521	PR	12712	34424
TR829	0.500	0.582	PR	551	34424
TR833	0.613	0.660	PR	1898	34424
TR837	0.432	0.537	DR	10541	34424
TR848	0.580	0.609	PR	8374	34424
TR854	0.582	0.664	NR	10691	34424
TR856	0.616	0.626	PR	2153	34424
TR857	0.328	0.417	PR	13203	34424
TR862	0.366	0.418	CMM	16176	60218
TR868	0.573	0.647	CME	20810	60218
TR869	0.289	0.305	CME	33886	60218
TR870	0.228	0.262	PR	42771	60218
TR872	0.568	0.642	CMM	10448	60218
TR879	0.633	0.634	CMM	5976	60218
TR891	0.757	0.784	CME	4368	60218
TR893	0.691	0.701	CME	18196	60218
TR921	0.484	0.513	CMM	26096	60218
TR928	0.430	0.415	CME	5264	60218
TR944	0.560	0.586	PR	20462	60218
TR945	0.412	0.436	CMM	47301	60218

APPENDIX D

Supplemental Material for Chapter 6 : ”Learning to Predict Improved Conformations of Proteins with Deep Recurrent Neural Networks”

Table D.2: Cross validation folds for the CASP dataset.

Traj. Name	Fold	# Snap.	# Traj.	I	N	D
TR759_dist_rst	0	12008	8	7401	1995	2612
TR759_no_rst	0	11208	8	4260	2231	4717
TR759_point_rst	0	11208	8	5517	1701	3990
TR782_no_rst	0	11208	8	30	171	11007
TR782_point_rst	0	11208	8	149	948	10111
TR810_dist_rst	0	12008	8	18	474	11516
TR810_no_rst	0	11208	8	55	690	10463
TR810_point_rst	0	11208	8	92	7104	4012
TR856_dist_rst	0	12008	8	0	1	12007
TR856_no_rst	0	11208	8	0	3	11205
TR856_point_rst	0	11208	8	0	267	10941
TR869_cm_excl	0	24505	5	18	3077	21410
TR869_cm_min	0	24505	5	3	882	23620
TR869_point_rst	0	11208	8	35	2599	8574
TR891_cm_excl	0	24505	5	4	1895	22606
TR891_cm_min	0	24505	5	28	2622	21855
TR891_point_rst	0	11208	8	12	2230	8966
TR283_dist_rst	1	12008	8	318	1074	10616

Table D.2: Cross validation folds for the CASP dataset.

Traj. Name	Fold	# Snap.	# Traj.	I	N	D
TR283_no_rst	1	11208	8	19	531	10658
TR283_point_rst	1	11208	8	0	8	11200
TR780_dist_rst	1	12008	8	4242	1940	5826
TR780_no_rst	1	11208	8	1167	1191	8850
TR780_point_rst	1	11208	8	5729	4651	828
TR837_dist_rst	1	12008	8	1268	886	9854
TR837_no_rst	1	11208	8	1329	1448	8431
TR837_point_rst	1	11208	8	727	1336	9145
TR854_dist_rst	1	12008	8	329	991	10688
TR854_no_rst	1	11208	8	376	1084	9748
TR854_point_rst	1	11208	8	2943	6376	1889
TR879_cm_excl	1	24505	5	0	11	24494
TR879_cm_min	1	24505	5	0	27	24478
TR879_point_rst	1	11208	8	0	40	11168
TR921_cm_excl	1	24505	5	17	6751	17737
TR921_cm_min	1	24505	5	27	9337	15141
TR921_point_rst	1	11208	8	13	6552	4643
TR217_dist_rst	2	12008	8	0	13	11995
TR217_no_rst	2	11208	8	0	15	11193
TR217_point_rst	2	11208	8	583	6864	3761
TR760_dist_rst	2	12008	8	0	8	12000
TR760_no_rst	2	11208	8	0	8	11200
TR760_point_rst	2	11208	8	381	5509	5318
TR786_dist_rst	2	12008	8	2	74	11932
TR786_no_rst	2	11208	8	1	60	11147
TR786_point_rst	2	11208	8	3451	4746	3011
TR816_dist_rst	2	12008	8	1017	488	10503
TR816_no_rst	2	11208	8	1598	1024	8586
TR816_point_rst	2	11208	8	1409	1754	8045
TR862_cm_excl	2	24505	5	422	3323	20760
TR862_cm_min	2	24505	5	2444	4205	17856
TR862_point_rst	2	11208	8	595	2900	7713
TR872_cm_excl	2	24505	5	2201	6844	15460
TR872_cm_min	2	24505	5	1488	4055	18962
TR872_point_rst	2	11208	8	767	3268	7173
TR762_dist_rst	3	12008	8	0	8	12000
TR762_no_rst	3	11208	8	0	8	11200
TR762_point_rst	3	11208	8	0	1656	9552
TR765_dist_rst	3	12008	8	11107	551	350
TR765_no_rst	3	11208	8	8259	963	1986
TR765_point_rst	3	11208	8	10719	425	64
TR828_dist_rst	3	12008	8	3	13	11992

Table D.2: Cross validation folds for the CASP dataset.

Traj. Name	Fold	# Snap.	# Traj.	I	N	D
TR828_no_rst	3	11208	8	1	15	11192
TR828_point_rst	3	11208	8	7	57	11144
TR833_dist_rst	3	12008	8	2	83	11923
TR833_no_rst	3	11208	8	1	64	11143
TR833_point_rst	3	11208	8	1232	4336	5640
TR928_cm_excl	3	24505	5	0	0	24505
TR928_cm_min	3	24505	5	0	0	24505
TR928_point_rst	3	11208	8	0	0	11208
TR945_cm_excl	3	24505	5	139	13368	10998
TR945_cm_min	3	24505	5	116	10678	13711
TR945_point_rst	3	11208	8	265	7400	3543
TR817_dist_rst	4	12008	8	843	958	10207
TR817_no_rst	4	11208	8	326	576	10306
TR817_point_rst	4	11208	8	73	6538	4597
TR821_dist_rst	4	12008	8	5115	1530	5363
TR821_no_rst	4	11208	8	5235	1458	4515
TR829_dist_rst	4	12008	8	84	425	11499
TR829_no_rst	4	11208	8	41	371	10796
TR829_point_rst	4	11208	8	3213	4337	3658
TR857_dist_rst	4	12008	8	1933	2781	7294
TR857_no_rst	4	11208	8	1734	1944	7530
TR857_point_rst	4	11208	8	1870	3345	5993
TR870_cm_excl	4	24505	5	17	343	24145
TR870_cm_min	4	24505	5	617	2950	20938
TR870_point_rst	4	11208	8	153	1718	9337
TR944_cm_excl	4	24505	5	86	2461	21958
TR944_cm_min	4	24505	5	578	3006	20921
TR944_point_rst	4	11208	8	59	2130	9019
TR769_dist_rst	5	12008	8	2004	1537	8467
TR769_no_rst	5	11208	8	3089	1880	6239
TR774_dist_rst	5	12008	8	1	46	11961
TR774_no_rst	5	11208	8	1	44	11163
TR774_point_rst	5	11208	8	6	1028	10174
TR792_dist_rst	5	12008	8	2815	2383	6810
TR792_no_rst	5	11208	8	3906	2620	4682
TR792_point_rst	5	11208	8	2811	4515	3882
TR795_dist_rst	5	12008	8	373	1367	10268
TR795_point_rst	5	11208	8	1956	7702	1550
TR848_dist_rst	5	12008	8	1	22	11985
TR848_no_rst	5	11208	8	2	53	11153
TR848_point_rst	5	11208	8	77	1541	9590
TR893_cm_excl	5	24505	5	0	121	24384

Table D.2: Cross validation folds for the CASP dataset.

Traj. Name	Fold	# Snap.	# Traj.	I	N	D
TR893_cm_min	5	24505	5	0	21	24484
TR893_point_rst	5	11208	8	0	45	11163
TR228_dist_rst	6	12008	8	3073	3563	5372
TR228_no_rst	6	11208	8	4852	2362	3994
TR228_point_rst	6	11208	8	5893	2875	2440
TR768_dist_rst	6	12008	8	50	591	11367
TR768_no_rst	6	11208	8	59	491	10658
TR768_point_rst	6	11208	8	1190	4161	5857
TR776_dist_rst	6	12008	8	0	20	11988
TR776_no_rst	6	11208	8	0	15	11193
TR776_point_rst	6	11208	8	312	7811	3085
TR783_dist_rst	6	12008	8	10	49	11949
TR783_no_rst	6	11208	8	0	55	11153
TR783_point_rst	6	11208	8	93	1640	9475
TR803_dist_rst	6	12008	8	215	1387	10406
TR803_no_rst	6	11208	8	8	351	10849
TR803_point_rst	6	11208	8	228	1006	9974
TR868_cm_excl	6	24505	5	329	542	23634
TR868_cm_min	6	24505	5	76	503	23926
TR868_point_rst	6	11208	8	592	1293	9323

Table D.3: CV performance all folds. Classes improved, no change and decreased are abbreviated with I, N and D respectively.

			Method			
Class	Fold	Metric	RNN	KNN	RF	LR
I	0	F1	0.039	0.059	0.001	0.000
I	0	Precision	0.310	0.078	0.083	0.000
I	0	Recall	0.021	0.048	0.001	0.000
I	1	F1	0.031	0.076	0.002	0.000
I	1	Precision	0.122	0.103	0.099	0.000
I	1	Recall	0.018	0.061	0.001	0.000
I	2	F1	0.060	0.085	0.001	0.000
I	2	Precision	0.138	0.095	0.207	0.000
I	2	Recall	0.039	0.077	0.001	0.000
I	3	F1	0.019	0.078	0.001	0.000
I	3	Precision	0.893	0.177	0.143	0.000
I	3	Recall	0.009	0.050	0.000	0.000
I	4	F1	0.128	0.095	0.003	0.000
I	4	Precision	0.454	0.146	0.347	0.000
I	4	Recall	0.074	0.071	0.002	0.000
I	5	F1	0.053	0.087	0.001	0.000

Table D.3: CB1 performance all folds. Classes improved, no change and decreased are abbreviated with I, N and D respectively.

			Method			
Class	**Fold**	**Metric**	**RNN**	**KNN**	**RF**	**LR**
I	5	Precision	0.551	0.125	0.034	0.000
I	5	Recall	0.028	0.067	0.001	0.000
I	6	F1	0.128	0.100	0.000	0.000
I	6	Precision	0.440	0.126	0.067	0.000
I	6	Recall	0.075	0.083	0.000	0.000
N	0	F1	0.138	0.202	0.116	0.117
N	0	Precision	0.269	0.222	0.275	0.295
N	0	Recall	0.093	0.185	0.074	0.073
N	1	F1	0.058	0.215	0.042	0.020
N	1	Precision	0.229	0.322	0.372	0.178
N	1	Recall	0.033	0.161	0.022	0.010
N	2	F1	0.164	0.260	0.096	0.025
N	2	Precision	0.393	0.365	0.504	0.481
N	2	Recall	0.103	0.202	0.053	0.013
N	3	F1	0.044	0.127	0.050	0.106
N	3	Precision	0.178	0.180	0.240	0.370
N	3	Recall	0.025	0.098	0.028	0.062
N	4	F1	0.124	0.229	0.109	0.074
N	4	Precision	0.214	0.297	0.499	0.366
N	4	Recall	0.088	0.186	0.061	0.041
N	5	F1	0.226	0.195	0.045	0.050
N	5	Precision	0.274	0.220	0.172	0.225
N	5	Recall	0.192	0.175	0.026	0.028
N	6	F1	0.226	0.262	0.121	0.088
N	6	Precision	0.324	0.277	0.390	0.322
N	6	Recall	0.173	0.249	0.072	0.051
D	0	F1	0.885	0.850	0.887	0.890
D	0	Precision	0.818	0.827	0.815	0.817
D	0	Recall	0.963	0.875	0.973	0.978
D	1	F1	0.856	0.840	0.861	0.858
D	1	Precision	0.764	0.785	0.760	0.757
D	1	Recall	0.973	0.904	0.993	0.990
D	2	F1	0.859	0.843	0.866	0.866
D	2	Precision	0.779	0.799	0.769	0.764
D	2	Recall	0.958	0.893	0.990	0.998
D	3	F1	0.837	0.804	0.835	0.838
D	3	Precision	0.728	0.731	0.725	0.730
D	3	Recall	0.983	0.892	0.985	0.983
D	4	F1	0.865	0.840	0.866	0.865
D	4	Precision	0.789	0.790	0.769	0.768

Table D.3: CV performance all folds. Classes improved, no change and decreased are abbreviated with I, N and D respectively.

			Method			
Class	**Fold**	**Metric**	**RNN**	**KNN**	**RF**	**LR**
D	4	Recall	0.957	0.895	0.991	0.991
D	5	F1	0.877	0.852	0.882	0.884
D	5	Precision	0.823	0.821	0.801	0.801
D	5	Recall	0.939	0.885	0.982	0.987
D	6	F1	0.887	0.857	0.890	0.888
D	6	Precision	0.830	0.837	0.811	0.808
D	6	Recall	0.952	0.877	0.986	0.986

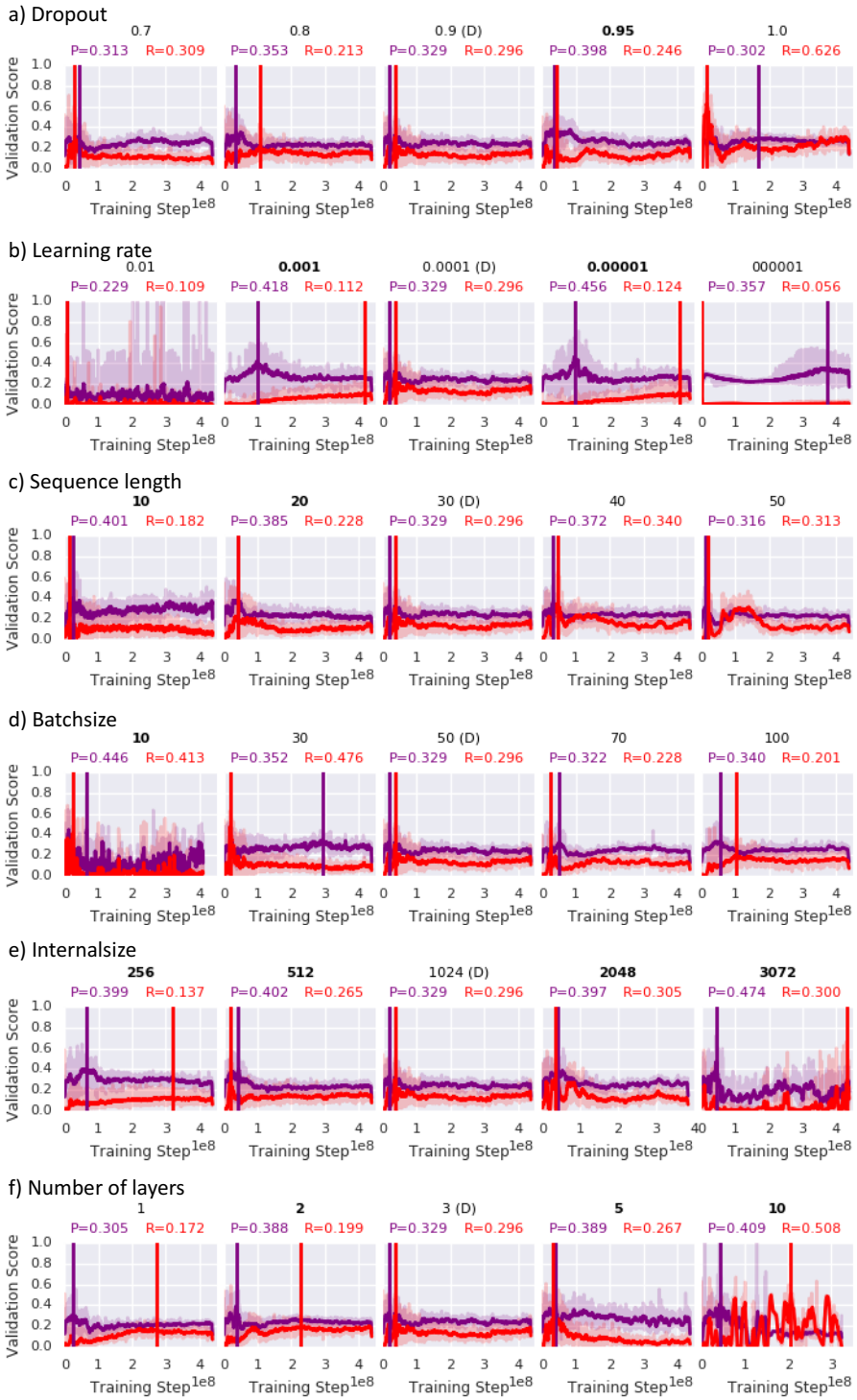

Figure D.1: Exploration of RNN hyper-parameters. The validation performance for recall (red) and precision (purple) of the no change class as a function of training steps for a) dropout, b) learning rate, c) sequence length d) batchsize, e) number of layers. The plot-title shows the tested parameter value where D in brackets indicates the default parameter and a bold value indicates an > 0.05 improvement in precision over the default parameter. The vertical lines in each sub-plot indicate the best running average with a window size of 30, scores are shown in the sub-plot header.

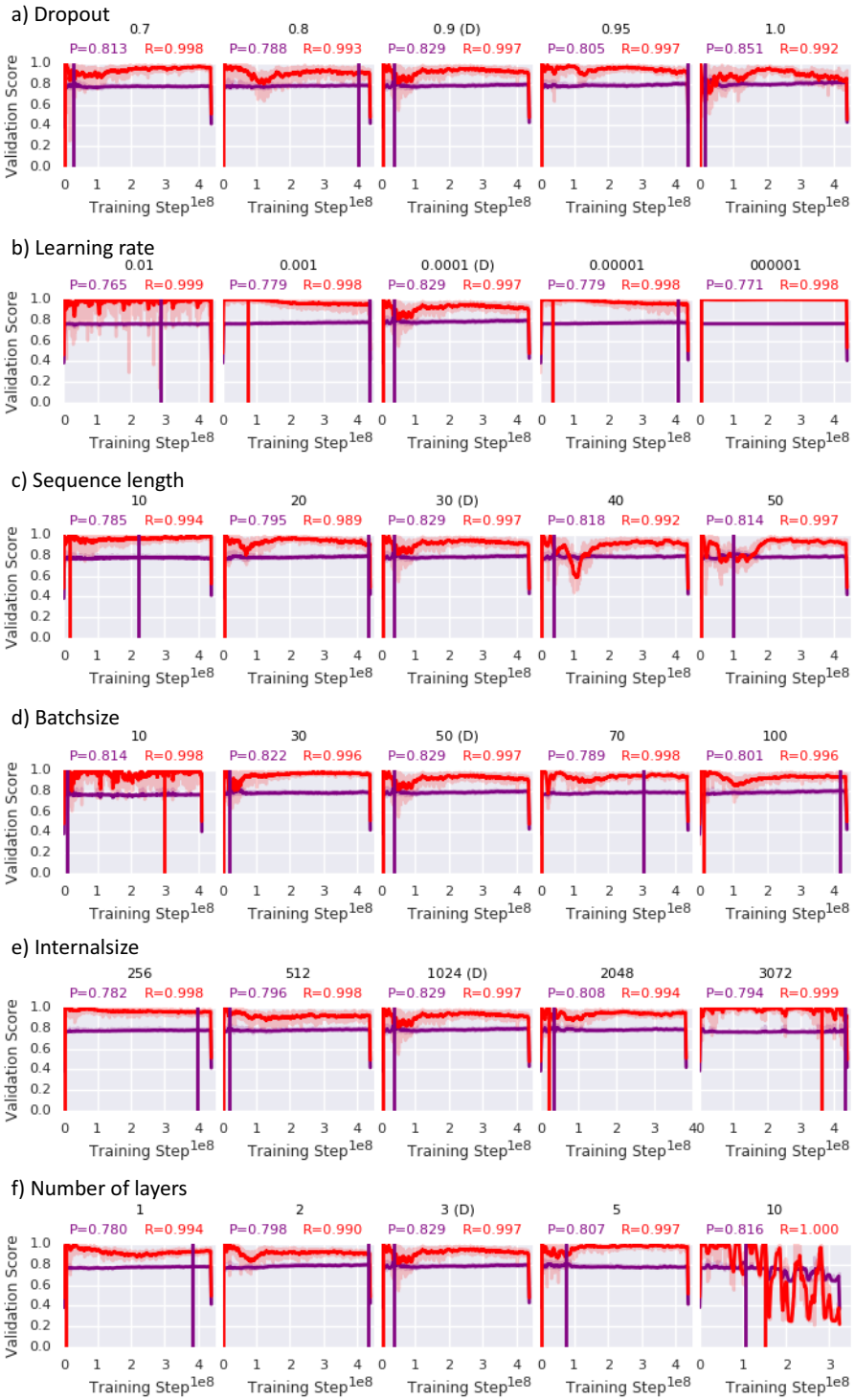

Figure D.2: Exploration of RNN hyper-parameters. The validation performance for recall (red) and precision (purple) of the decreased class as a function of training steps for a) dropout, b) learning rate, c) sequence length d) batchsize, e) number of layers. The plot-title shows the tested parameter value where D in brackets indicates the default parameter and a bold value indicates an > 0.05 improvement in precision over the default parameter. The vertical lines in each sub-plot indicate the best running average with a window size of 30, scores are shown in the sub-plot header.

APPENDIX D: SUPPLEMENTAL MATERIAL FOR CHAPTER 6 : "LEARNING TO PREDICT IMPROVED CONFORMATIONS OF PROTEINS WITH DEEP RECURRENT NEURAL NETWORKS"

Table D.1: RNN features. Table lists the feature name, the descriptor that produced the feature and the reference for the descriptor.

Feature	Descriptor	Reference
N_DDFIRESUM	DFIRE	(Liu et al., 2004)
N_DDFIRETERM1	DFIRE	(Liu et al., 2004)
N_DDFIRETERM2	DFIRE	(Liu et al., 2004)
N_DDFIRETERM3	DFIRE	(Liu et al., 2004)
N_DDFIRETERM4	DFIRE	(Liu et al., 2004)
N_DOPE	DOPE	(Shen and Sali, 2006)
N_DOPE_HR	DOPE	(Shen and Sali, 2006)
N_RMSD_SM	RMSD to starting model	see Section 2.6.1
N_GDTTS_SM	GDTTS to staring model	see Section 2.6.1
N_DOOP	DOOP	(Chae et al., 2015)
N_CALRW	calRW	(Zhang and Zhang, 2010)
N_CALRWP	calRWplus	(Zhang and Zhang, 2010)
N_GOAP	GOAP	(Zhou and Skolnick, 2011)
N_GOAPAG	GOAP	(Zhou and Skolnick, 2011)
N_BOND	Modeller	(Eswar et al., 2008)
N_ANGLE	Modeller	(Eswar et al., 2008)
N_DIHEDRAL	Modeller	(Eswar et al., 2008)
N_IMPROPER	Modeller	(Eswar et al., 2008)
N_MOLPDF	Mol. PDF	(Eswar et al., 2008)

Bibliography

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2016). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. arXiv:1603.04467.

Agius, R., Torchala, M., Moal, I. H., Fernández-Recio, J., and Bates, P. A. (2013). Characterizing changes in the rate of protein-protein dissociation upon interface mutation using hotspot energy and organization. *PLoS Comput Biol*, 9(9):e1003216.

Airaksinen, M. S., Titievsky, A., and Saarma, M. (1999). GDNF Family Neurotrophic Factor Signaling: Four Masters, One Servant? *Molecular and Cellular Neuroscience*, 13(5):313–325.

Amoresano, A., Incoronato, M., Monti, G., Pucci, P., De Franciscis, V., and Cerchia, L. (2005). Direct interactions among Ret, GDNF and GFRA1 molecules reveal new insights into the assembly of a functional three-protein complex. *Cellular Signalling*, 17(6):717–727.

Anders, J., Kjær, S., and Ibáñez, C. F. (2001). Molecular Modeling of the Extracellular Domain of the RET Receptor Tyrosine Kinase Reveals Multiple Cadherin-like Domains and a Calcium-binding Site. *Journal of Biological Chemistry*, 276(38):35808–35817.

Andrusier, N., Nussinov, R., and Wolfson, H. J. (2007). FireDock: fast interaction refinement in molecular docking. *Proteins: Structure, Function, and Bioinformatics*, 69(1):139–159, 0605018.

Angrist, M., Jing, S., Bolk, S., Bentley, K., Nallasamy, S., Halushka, M., Fox, G. M., and Chakravarti, A. (1998). Human GFRA1: cloning, mapping, genomic structure, and evaluation as a candidate gene for Hirschsprung disease susceptibility. *Genomics*, 48(3):354–362.

Arighi, E., Borrello, M. G., and Sariola, H. (2005). RET tyrosine kinase signaling in development and cancer. *Cytokine & Growth Factor Reviews*, 16(4-5):441–467.

Asimov, I. (1950). *I , Robot*. Gnome Press.

Baker, D. and Sali, A. (2001). Protein structure prediction and structural genomics. *Science*, 294(5540):93–96.

Barducci, A., Bonomi, M., and Parrinello, M. (2011). Metadynamics. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(5):826–843.

Barducci, A., Bussi, G., and Parrinello, M. (2008). Well-tempered metadynamics: A smoothly converging and tunable free-energy method. *Physical Review Letters*, 100(2):020603, 0803.3861.

Bengio, Y., De Mori, R., Flammia, G., and Kompe, R. (1992). Phonetically motivated acoustic parameters for continuous speech recognition using artificial neural networks. *Speech Communication*, 11(2-3):261–271.

Bentley, J. L. (1975). Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517.

Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., and Haak, J. R. (1984). Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81(8):3684–3690.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The protein data bank. *Nucleic acids research*, 28(1):235–242.

Berteotti, A., Cavalli, A., Branduardi, D., Gervasio, F. L., Recanatini, M., and Parrinello, M. (2009). Protein conformational transitions: The closure mechanism of a kinase explored by atomistic simulations. *Journal of the American Chemical Society*, 131(1):244–250.

Besset, V., Scott, R. P., and Ibáñez, C. F. (2000). Signaling complexes and protein-protein interactions involved in the activation of the Ras and phosphatidylinositol 3-kinase pathways by the c-Ret receptor tyrosine kinase. *Journal of Biological Chemistry*, 275(50):39159–39166.

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., Kiefer, F., Cassarino, T. G., Bertoni, M., Bordoli, L., and Schwede, T. (2014). SWISS-MODEL: Modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Research*, 42(W1).

Bishop, C. (2006). *Pattern recognition and machine learning*, volume 4. Springer, 0-387-31073-8.

Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L. D., Monfort, M., Muller, U., Zhang, J., Zhang, X., Zhao, J., and Zieba, K. (2016). End to End Learning for Self-Driving Cars. arXiv:1604.07316.

Bonanomi, D., Chivatakarn, O., Bai, G., Abdesselem, H., Lettieri, K., Marquardt, T., Pierchala, B. A., and Pfaff, S. L. (2012). Ret is a multifunctional coreceptor that integrates diffusible- and contact-axon guidance signals. *Cell*, 148(3):568–582.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Brooks, B. R., Brooks, C. L., MacKerell, A. D., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., and Boresch, S. (2009). CHARMM: the biomolecular simulation program. *Journal of computational chemistry*, 30(10):1545–1614.

Brown, P. F., Cocke, J., Della-Pietra, S. a., Della-Pietra, V. J., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Rossin, P. (1990). A statistical approach to machine translation. *Computational Linguistics*, 16(2):76–85.

Bussi, G., Donadio, D., and Parrinello, M. (2007). Canonical sampling through velocity rescaling. *Journal of Chemical Physics*, 126(1), arXiv:0803.4060v1.

CAFA (2016). An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biology*, 1601.00891.

Caflisch, R. E. (1998). *Monte Carlo and quasi-Monte Carlo methods*, volume 7. Cambridge University Press.

Chae, M. H., Krull, F., and Knapp, E. W. (2015). Optimized distance-dependent atom-pair-based potential DOOP for protein structure prediction. *Proteins: Structure, Function and Bioinformatics*, 83(5):881–890.

Chatterjee, S., Debenedetti, P. G., Stillinger, F. H., and Lynden-Bell, R. M. (2008). A computational investigation of thermodynamics, structure, dynamics and solvation behavior in modified water models. *Journal of Chemical Physics*, 128(12):124511.

Chaudhury, S., Berrondo, M., Weitzner, B. D., Muthu, P., Bergman, H., and Gray, J. J. (2011). Benchmarking and analysis of protein docking performance in Rosetta v3.2. *PLoS ONE*, 6(8).

Chaudhury, S., Lyskov, S., and Gray, J. J. (2010). PyRosetta: A script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics*, 26(5):689–691.

Chellapilla, K., Puri, S., and Simard, P. (2006). High Performance Convolutional Neural Networks for Document Processing. In *Tenth International Workshop on Frontiers in Handwriting Recognition*, pages 1–6.

Chen, R., Li, L., and Weng, Z. (2003). ZDOCK: an initialstage proteindocking algorithm. *Proteins: Structure, Function, and Bioinformatics*, 52(1):80–87.

Chen, Y.-h., Hu, L., Punta, M., Bruni, R., Hillerich, B., Kloss, B., Rost, B., Love, J., Siegelbaum, S. A., and Hendrickson, W. A. (2010). Homologue structure of the SLAC1 anion channel for closing stomata in leaves. *Nature*, 467(7319):1074–1080.

Chen, Z., Li, Y., Chen, E., Hall, D. L., Darke, P. L., Culberson, C., Shafer, J. A., and Kuo, L. C. (1994). Crystal Structure at 1.9-Angstrom Resolution of Human Immunodeficiency Virus (HIV) II Protease Complexed with L-735,524, an Orally Bioavailable Inhibitor of the HIV Proteases. *Journal of Biological Chemistry*, 269(42):26344–26348.

Cheng, T. M. K., Blundell, T. L., and Fernandez-Recio, J. (2007). PyDock: Electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. *Proteins: Structure, Function and Genetics*, 68(2):503–515, 0605018.

Chi, X., Michos, O., Shakya, R., Riccio, P., Enomoto, H., Licht, J. D., Asai, N., Takahashi, M., Ohgami, N., Kato, M., Mendelsohn, C., and Costantini, F. (2009). Ret-Dependent Cell Rearrangements in the Wolffian Duct Epithelium Initiate Ureteric Bud Morphogenesis. *Developmental Cell*, 17(2):199–209.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. 1406.1078.

Chuang, G.-Y., Kozakov, D., Brenke, R., Comeau, S. R., and Vajda, S. (2008). DARS (Decoys As the Reference State) Potentials for Protein-Protein Docking. *Biophysical Journal*, 95(9):4217–4227.

Cirean, D. C., Meier, U., Maria, L., and Ch Urgen Schmidhuber, G. L. (2010). Deep, Big, Simple Neural Nets for Handwritten Digit Recognition. *Neural Computation*, arXiv:1003.0358.

Coates, A., Huval, B., Wang, T., Wu, D., and Ng, A. Y. (2013). Deep learning with COTS HPC systems. In *Proceedings of The 30th International Conference on Machine Learning*.

Coates, A. and Ng, A. (2011). The importance of encoding versus training with sparse coding and vector quantization. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 921—-928.

Cockburn, J. G., Richardson, D. S., Gujral, T. S., and Mulligan, L. M. (2010). RET-mediated cell adhesion and migration require multiple integrin subunits. *Journal of Clinical Endocrinology and Metabolism*, 95(11).

Comeau, S. R., Gatchell, D. W., Vajda, S., and Camacho, C. J. (2004). ClusPro: An automated docking and discrimination method for the prediction of protein complexes. *Bioinformatics*, 20(1):45–50.

Coulpier, M., Anders, J., and Ibáñez, C. F. (2002). Coordinated activation of autophosphorylation sites in the RET receptor tyrosine kinase: Importance of tyrosine 1062 for GDNF mediated neuronal differentiation and survival. *Journal of Biological Chemistry*, 277(3):1991–1999.

Cozzetto, D., Buchan, D. W., Bryson, K., and Jones, D. T. (2013). Protein function prediction by massive integration of evolutionary analyses and multiple data sources. *BMC Bioinformatics*, 14(Suppl 3):S1.

Cunningham, P. and Delany, S. J. (2007). K -Nearest Neighbour Classifiers. *Multiple Classifier Systems*, pages 1–17.

Darden, T., York, D., and Pedersen, L. (1993). Particle mesh Ewald: An N log(N) method for Ewald sums in large systems. *The Journal of Chemical Physics*, 98(12):10089–10092, 9807099.

Daura, X., Gademann, K., Jaun, B., Seebach, D., van Gunsteren, W. F., and Mark, A. E. (1999). Peptide folding: when simulation meets experiment. *Angewandte Chemie International Edition*, 38(12):236–240.

Davis, T. K., Hoshi, M., and Jain, S. (2014). To bud or not to bud: The RET perspective in CAKUT. *Pediatric Nephrology*, 29(4):597–608.

de Juan, D., Pazos, F., and Valencia, A. (2013). Emerging methods in protein co-evolution. *Nature Reviews Genetics*, 14(4):249–261.

De Vita, G., Melillo, R. M., Carlomagno, F., Visconti, R., Castellone, M. D., Bellacosa, A., Billaud, M., Fusco, A., Tsichlis, P. N., and Santoro, M. (2000). Tyrosine 1062 of RET-MEN2A mediates activation of Akt (protein kinase B) and mitogen-activated protein kinase pathways leading to PC12 cell survival. *Cancer Research*, 60(19):3727–3731.

Deng, J., Berg, A. C., Li, K., and Fei-Fei, L. (2010). What does classifying more than 10,000 image categories tell us? In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 6315 LNCS, pages 71–84.

Dill, K. A. and MacCallum, J. L. (2012). The Protein-Folding Problem, 50 Years On. *Science*, 338(6110):1042–1046.

Dominguez, C., Boelens, R., and Bonvin, A. M. (2003). HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society*, 125(7):1731–1737.

Donis-keller, H., Dou, S., Chi, D., Carlson, K. M., Toshima, K., Lairmore, T. C., Howe, J. R., Moley, J. F., Goodfellow, P., and Wells, S. A. (1993). Mutations in the RET proto-oncogene are associated with MEN 2a and FMTC. *Human Molecular Genetics*, 2(7):851–856.

Durbec, P. L., Larsson-Blomberg, L. B., Schuchardt, A., Costantini, F., and Pachnis, V. (1996). Common origin and developmental dependence on c-ret of subsets of enteric and sympathetic neuroblasts. *Development (Cambridge, England)*, 122(1):349–58.

Eisenstein, M. and Katchalski-Katzir, E. (2004). On proteins, grids, correlations, and docking. *Comptes Rendus - Biologies*, 327(5):409–420.

Esposito, C. L., D'Alessio, A., de Franciscis, V., and Cerchia, L. (2008). A cross-talk between TrkB and ret tyrosine kinase receptors mediates neuroblastoma cells differentiation. *PLoS ONE*, 3(2).

Eswar, N., Eramian, D., Webb, B., Shen, M.-Y., and Sali, A. (2008). Protein Structure Modeling with MODELLER. In *Protein Structure Prediction*, pages 145–159. arXiv:1011.1669v3.

Fehske, H., Schneider, R., and Weisse, A. (2007). *Computational Many-Particle Physics*. Springer.

Feig, M. (2017). Computational protein structure refinement: Almost there, yet still so far to go. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 7(3):e1307.

Feig, M. and Mirjalili, V. (2016). Protein structure refinement via molecular-dynamics simulations: What works and what does not? *Proteins: Structure, Function, and Bioinformatics*, 84(S1):282–292.

Feliu, E., Aloy, P., and Oliva, B. (2011). On the analysis of protein-protein interactions via knowledge-based potentials for the prediction of protein-protein docking. *Protein Science*, 20(3):529–541.

Feng, Y., Kloczkowski, A., and Jernigan, R. L. (2010). Potentials 'R'Us web-server for protein energy estimations with coarse-grained knowledge-based potentials. *BMC Bioinformatics*, 11(1):92.

Ferguson, K. M. (2008). Structure-Based View of Epidermal Growth Factor Receptor Regulation. *Annual Review of Biophysics*, 37(1):353–373.

Fernández-Recio, J., Totrov, M., and Abagyan, R. (2003). ICM-DISCO docking by global energy optimization with fully flexible side-chains. *Proteins: Structure, Function and Genetics*, 52(1):113–117.

FernandezRecio, J., Totrov, M., Skorodumov, C., and Abagyan, R. (2005). Optimal docking area: a new method for predicting proteinprotein interaction sites. *Proteins: Structure, Function, and Bioinformatics*, 58(1):134–143.

Filippakopoulos, P., Qi, J., Picaud, S., Shen, Y., Smith, W. B., Fedorov, O., Morse, E. M., Keates, T., Hickman, T. T., Felletar, I., Philpott, M., Munro, S., McKeown, M. R., Wang, Y., Christie, A. L., West, N., Cameron, M. J., Schwartz, B.,

Heightman, T. D., La Thangue, N., French, C. A., Wiest, O., Kung, A. L., Knapp, S., and Bradner, J. E. (2010). Selective inhibition of BET bromodomains. *Nature*, 468(7327):1067–1073.

Fink, F., Hochrein, J., Wolowski, V., Merkl, R., and Gronwald, W. (2011). PROCOS: Computational analysis of proteinprotein complexes. *Journal of computational chemistry*, 32(12):2575–2586.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, arXiv:1011.1669v3.

Foda, Z. H., Shan, Y., Kim, E. T., Shaw, D. E., and Seeliger, M. A. (2015). A dynamically coupled allosteric network underlies binding cooperativity in Src kinase. *Nature Communications*, 6:5939.

French, C. A., Miyoshi, I., Kubonishi, I., Grier, H. E., Perez-Atayde, A. R., and Fletcher, J. A. (2003). BRD4-NUT fusion oncogene: A novel mechanism in aggressive carcinoma. *Cancer Research*, 63(2):304–307.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, arXiv:1011.1669v3.

Gao, Y., Cao, E., Julius, D., and Cheng, Y. (2016). TRPV1 structures in nanodiscs reveal mechanisms of ligand and lipid action. *Nature*, 534(7607):347–351, 15334406.

Garson, J. G. (1900). The metric system of identification of criminals, as used in in Great Britain and Ireland. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 30(2):177–227.

Gattei, V., Celetti, A., Cerrato, A., Degan, M., De Iuliis, A., Rossi, F. M., Chiappetta, G., Consales, C., Improta, S., Zagonel, V., Aldinucci, D., Agosti, V., Santoro, M., Vecchio, G., Pinto, A., and Grieco, M. (1997). Expression of the RET receptor tyrosine kinase and GDNFR-alpha in normal and leukemic human hematopoietic cells and stromal cells of the bone marrow microenvironment. *Blood*, 89(8):2925–37.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, 63(1):3–42.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press, arXiv:1011.1669v3.

Goodfellow, I., Pouget-Abadie, J., and Mirza, M. (2014). Generative Adversarial Networks. pages 1–9, arXiv:1406.2661v1.

Gujral, T. S., Singh, V. K., Jia, Z., and Mulligan, L. M. (2006). Molecular mechanisms of RET receptor-mediated oncogenesis in multiple endocrine neoplasia 2B. *Cancer Research*, 66(22):10741–10749.

Hamp, T. and Rost, B. (2012). Alternative protein-protein interfaces are frequent exceptions. *PLoS Comput Biol*, 8(8):e1002623.

Hayashi, H., Ichihara, M., Iwashita, T., Murakami, H., Shimono, Y., Kawai, K., Kurokawa, K., Murakumo, Y., Imai, T., Funahashi, H., Nakao, A., and Takahashi, M. (2000). Characterization of intracellular signals via tyrosine 1062 in RET activated by glial cell line-derived neurotrophic factor. *Oncogene*, 19(39):4469–4475.

Hess, B., Kutzner, C., Van Der Spoel, D., and Lindahl, E. (2008). GRGMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of Chemical Theory and Computation*, 4(3):435–447.

Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006). A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, 18(7):1527–1554, arXiv:1111.6189v1.

Hochreiter, S. and Schmidhuber, J. (1997). LONG SHORT-TERM MEMORY. *Neural Computation*, 9(8):1735–1780, 1206.2944.

Hockney, R. W. and Eastwood, J. W. (1988). *Computer Simulation Using Particles*, volume 25. SIAM Review.

Hopf, T. A., Schärfe, C. P. I., Rodrigues, J. P. G. L. M., Green, A. G., Kohlbacher, O., Sander, C., Bonvin, A. M. J. J., and Marks, D. S. (2014). Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife*, 3, arXiv:1405.0929.

Huang, Y. J., Mao, B., Aramini, J. M., and Montelione, G. T. (2014). Assessment of template-based protein structure predictions in CASP10. *Proteins: Structure, Function and Bioinformatics*, 82(SUPPL.2):43–56.

Jaeger, H. (2004). Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication. *Science*, 304(5667):78–80, arXiv:1011.1669v3.

Janeway, C. A., Travers, P., Walport, M., and Shlomchik, M. (2001). *Immunobiology: The Immune System In Health And Disease*. Garland Science.

Janin, J. (2010). Proteinprotein docking tested in blind predictions: the CAPRI experiment. *Molecular BioSystems*, 6(12):2351–2362.

Jarrett, K., Kavukcuoglu, K., Ranzato, M., and LeCun, Y. (2009). What is the best multi-stage architecture for object recognition? In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2146–2153.

Jia Deng, Wei Dong, Socher, R., Li-Jia Li, Kai Li, and Li Fei-Fei (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '02*, page 133.

Jones, D. T., Buchan, D. W. A., Cozzetto, D., and Pontil, M. (2012). PSICOV: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, 28(2):184–190.

Jones, D. T., Singh, T., Kosciolek, T., and Tetchner, S. (2015). MetaPSICOV: Combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics*, 31(7):999–1006.

Jones, J. E. (1924). On the Determination of Molecular Fields. I. From the Variation of the Viscosity of a Gas with Temperature. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 106(738):441–462.

Jones, S. and Thornton, J. M. (1996). Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences*, 93(1):13–20.

Joo, K., Joung, I., Lee, S. Y., Kim, J. Y., Cheng, Q., Manavalan, B., Joung, J. Y., Heo, S., Lee, J., Nam, M., Lee, I. H., Lee, S. J., and Lee, J. (2015). Template based protein structure modeling by global optimization in CASP11. *Proteins: Structure, Function and Bioinformatics*, 84(S1):221–232.

Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2):926–935, arXiv:1011.1669v3.

Jura, N., Endres, N. F., Engel, K., Deindl, S., Das, R., Lamers, M. H., Wemmer, D. E., Zhang, X., and Kuriyan, J. (2009). Mechanism for activation of the EGF receptor catalytic domain by the juxtamembrane segment. *Cell*, 137(7):1293–307.

Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogenbonded and geometrical features. *Biopolymers*, 22(12):2577–2637.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Li, F. F. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1725–1732. arXiv:1412.0767.

Karplus, K. (2009). SAM-T08, HMM-based protein structure prediction. *Nucleic Acids Research*, 37(SUPPL. 2):W492–7.

Kawamoto, Y., Takeda, K., Okuno, Y., Yamakawa, Y., Ito, Y., Taguchi, R., Kato, M., Suzuki, H., Takahashi, M., and Nakashima, I. (2004). Identification of RET Autophosphorylation Sites by Mass Spectrometry. *Journal of Biological Chemistry*, 279(14):14213–14224.

Kendrew, J. C., Bodo, G., Dintzs, H. M., Parrish, R. G., Wyckoff, H., and Phillips, D. C. (1958). A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. *Nature*, 181(4610):662–666, arXiv:1011.1669v3.

Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., Duesbury, M., Dumousseau, M., Feuermann, M., Hinz, U., Jandrasits, C., Jimenez, R. C., Khadake, J., Mahadevan, U., Masson, P., Pedruzzi, I., Pfeiffenberger, E., Porras, P., Raghunath, A., Roechert, B., Orchard, S., and Hermjakob, H. (2012). The IntAct molecular interaction database in 2012. *Nucleic Acids Research*, 40(D1).

Kim, D. E., Dimaio, F., Yu-Ruei Wang, R., Song, Y., and Baker, D. (2014). One contact for every twelve residues allows robust and accurate topology-level protein structure modeling. *Proteins: Structure, Function and Bioinformatics*, 82(SUPPL.2):208–218, 9605103.

Kinch, L. N., Li, W., Monastyrskyy, B., Kryshtafovych, A., and Grishin, N. V. (2016). Assessment of CASP11 contact-assisted predictions. *Proteins: Structure, Function and Bioinformatics*, 84:164–180.

Kingma, D. P. and Ba, J. (2014). Adam: A Method for Stochastic Optimization. arXiv:1412.6980.

Kjaer, S., Hanrahan, S., Totty, N., and McDonald, N. Q. (2010). Mammal-restricted elements predispose human RET to folding impairment by HSCR mutations. *Nature structural & molecular biology*, 17(6):726–731.

Knowles, P. P., Murray-Rust, J., Kjær, S., Scott, R. P., Hanrahan, S., Santoro, M., Ibáñez, C. F., and McDonald, N. Q. (2006). Structure and chemical inhibition of the RET tyrosine kinase domain. *Journal of Biological Chemistry*, 281(44):33577–33587.

Krivov, G. G., Shapovalov, M. V., and Dunbrack, R. L. (2009). Improved prediction of protein sidechain conformations with SCWRL4. *Proteins: Structure, Function, and Bioinformatics*, 77(4):778–795.

Krizhevsky, A., Hinton, G. E., and Sutskever, I. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In *Adv. Neural Inf. Process. Syst.* arXiv:1102.0183.

Krogsgaard, M., Li, Q.-j., Sumen, C., Huppa, J. B., Huse, M., and Davis, M. M. (2005). Agonist/endogenous peptideMHC heterodimers drive T cell activation and sensitivity. *Nature*, 434(7030):238–243.

Kryshtafovych, A., Monastyrskyy, B., Fidelis, K., Schwede, T., and Tramontano, A. (2017). Assessment of model accuracy estimations in CASP12. *Proteins: Structure, Function, and Bioinformatics*.

Kryshtafovych, A., Venclovas, Č., Fidelis, K., and Moult, J. (2005). Progress over the first decade of CASP experiments. *Proteins: Structure, Function and Genetics*, 61(SUPPL. 7):225–236.

Kubat, M. and Matwin, S. (1997). Addressing the curse of imbalanced training sets: one-sided selection. In *ICML*, volume 97, pages 179–186.

Kuroda, D. and Gray, J. J. (2016). Pushing the Backbone in Protein-Protein Docking. *Structure*, 24(10):1821–1829.

Lan, L., Djuric, N., Guo, Y., and Vucetic, S. (2013). MS-kNN: protein function prediction by integrating multiple data sources. *BMC bioinformatics*, 14 Suppl 3(Suppl 3):S8.

Le, Q. V., Ranzato, M., Monga, R., Devin, M., Chen, K., Corrado, G. S., Dean, J., and Ng, A. Y. (2011). Building high-level features using large scale unsupervised learning. In *International Conference in Machine Learning*.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2323, arXiv:1102.0183.

Lensink, M. F. and Wodak, S. J. (2013). Docking, scoring, and affinity prediction in CAPRI. *Proteins: Structure, Function, and Bioinformatics*, 81(12):2082–2095.

Lensink, M. F. and Wodak, S. J. (2014). Score_set: A CAPRI benchmark for scoring protein complexes. *Proteins: Structure, Function, and Bioinformatics*, 82(11):3163–3169.

Levinthal, C. (1968). Are there pathways for protein folding? *Journal de Chimie Physique*, 65:44–45.

Levinthal, C. (1969). How to fold graciously. *Mössbauer Spectroscopy in Biological Systems Proceedings*, 24(41):22–24.

Li, L., Umbach, D. M., Terry, P., and Taylor, J. A. (2004). Application of the GA/KNN method to SELDI proteomics data. *Bioinformatics*, 20(10):1638–1640.

Li, S. C., Bu, D., Xu, J., and Li, M. (2008). Fragment-HMM: A new approach to protein structure prediction. *Protein Science*, 17(11):1925–1934.

Liao, J. and Chin, K. (2007). Logistic regression for disease classification using microarray data: model selection in a large p and small n case. *Bioinformatics*, 23(15):1945–1951.

Lin, H. W. and Tegmark, M. (2016). Why does deep learning work so well? 02139, arXiv:1608.08225v1.

Lin, L., Doherty, D., Lile, J., Bektesh, S., and Collins, F. (1993). GDNF: a glial cell line-derived neurotrophic factor for midbrain dopaminergic neurons. *Science*, 260(5111):1130–1132.

Lindorff-Larsen, K., Piana, S., Dror, R. O., and Shaw, D. E. (2011). How Fast-Folding Proteins Fold. *Science*, 334(6055).

Liu, S. and Vakser, I. A. (2011). DECK: Distance and environment-dependent, coarse-grained, knowledge-based potentials for protein-protein docking. *BMC bioinformatics*, 12(1):280.

Liu, S., Zhang, C., Zhou, H., and Zhou, Y. (2004). A physical reference state unifies the structurederived potential of mean force for protein folding and binding. *Proteins: Structure, Function, and Bioinformatics*, 56(1):93–101.

Liu, T., Moore, A. W., and Gray, A. (2006). New Algorithms for Efficient High-Dimensional Nonparametric Classification. *Journal of Machine Learning Research*, 7:1135–1158.

Lovell, S. C., Davis, I. W., Arendall, W. B., De Bakker, P. I., Word, J. M., Prisant, M. G., Richardson, J. S., and Richardson, D. C. (2003). Structure validation by Cα geometry: ϕ,ψ and Cβ deviation. *Proteins: Structure, Function and Genetics*, 50(3):437–450.

Lu, H., Lu, L., and Skolnick, J. (2003). Development of unified statistical potentials describing protein-protein interactions. *Biophysical journal*, 84(March):1895–1901.

Lu, M., Dousis, A. D., and Ma, J. (2008). OPUS-PSP: An Orientation-dependent Statistical All-atom Potential Derived from Side-chain Packing. *Journal of Molecular Biology*, 376(1):288–301.

Lyskov, S. and Gray, J. J. (2008). The RosettaDock server for local protein-protein docking. *Nucleic acids research*, 36(Web Server issue):W233–W238.

Ma, J., Wang, S., Zhao, F., and Xu, J. (2013). Protein threading using context-specific alignment potential. In *Bioinformatics*, volume 29, pages i257–65. Oxford University Press.

Maccallum, J. L., Pérez, A., Schnieders, M. J., Hua, L., Jacobson, M. P., and Dill, K. A. (2011). Assessment of protein structure refinement in CASP9. *Proteins: Structure, Function and Bioinformatics*, 79(SUPPL. 10):74–90.

Manavalan, B. and Lee, J. (2017). SVMQA: supportvector-machine-based protein single-model quality assessment. *Bioinformatics*, 33(16):2496–2503.

Mandell, J. G., Roberts, V. A., Pique, M. E., Kotlovyi, V., Mitchell, J. C., Nelson, E., Tsigelny, I., and Ten Eyck, L. F. (2001). Protein docking using continuum electrostatics and geometric fit. *Protein engineering*, 14(2):105–113.

Mariani, V., Kiefer, F., Schmidt, T., Haas, J., and Schwede, T. (2011). Assessment of template based protein structure predictions in CASP9. *Proteins: Structure, Function and Bioinformatics*, 79(SUPPL. 10):37–58.

Marillet, S., Boudinot, P., and Cazals, F. (2016). High-resolution crystal structures leverage protein binding affinity predictions. *Proteins: Structure, Function and Bioinformatics*, 84(1):9–20.

Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R., and Sander, C. (2011). Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE*, 6(12), arXiv:1110.5091.

Marsh, J. A. and Teichmann, S. A. (2015). Structure, Dynamics, Assembly, and Evolution of Protein Complexes. *Annual Review of Biochemistry*, 84(1):551–575.

Meier, A. and Söding, J. (2015). Automatic Prediction of Protein 3D Structures by Probabilistic Multi-template Homology Modeling. *PLoS Computational Biology*, 11(10).

Méndez, R., Leplae, R., De Maria, L., and Wodak, S. J. (2003). Assessment of blind predictions of protein-protein interactions: current status of docking methods. *Proteins*, 52(1):51–67.

Méndez, R., Leplae, R., Lensink, M. F., and Wodak, S. J. (2005). Assessment of CAPRI predictions in rounds 35 shows progress in docking procedures. *Proteins: Structure, Function, and Bioinformatics*, 60(2):150–169.

Miller, A. N. and Long, S. B. (2012). Crystal Structure of the Human Two-Pore Domain Potassium Channel K2P1. *Science*, 335(6067):432–436.

Mintseris, J., Pierce, B., Wiehe, K., Anderson, R., Chen, R., and Weng, Z. (2007). Integrating statistical pair potentials into protein complex prediction. *Proteins: Structure, Function and Genetics*, 69(3):511–520, 0605018.

Mirjalili, V., Noyes, K., and Feig, M. (2014). Physics-based protein structure refinement through multiple molecular dynamics trajectories and structure averaging. *Proteins: Structure, Function and Bioinformatics*, 82(SUPPL.2):196–207.

Mitra, P. and Pal, D. (2010). New measures for estimating surface complementarity and packing at protein-protein interfaces. *FEBS Letters*, 584(6):1163–1168.

Miyoshi, I., Aster, J. C., Kubonishi, I., Kroll, T. G., Dal Cin, P., Vargas, S. O., Perez-Atayde, A. R., and Fletcher, J. A. (2001). BRD4 bromodomain gene rearrangement in aggressive carcinoma with translocation t(15;19). *American Journal of Pathology*, 159(6):1987–1992.

Moal, I. H., Agius, R., and Bates, P. A. (2011). Proteinprotein binding affinity prediction on a diverse set of structures. *Bioinformatics*, 27(21):3002–3009.

Moal, I. H., Barradas-Bautista, D., Jiménez-García, B., Torchala, M., Van Der Velde, A., Vreven, T., Weng, Z., Bates, P. A., and Fernández-Recio, J. (2017). IRaPPA: Information retrieval based integration of biophysical models for protein assembly selection. *Bioinformatics*, 33(12):1806–1813.

Moal, I. H. and Bates, P. A. (2010). SwarmDock and the use of normal modes in protein-protein docking. *International journal of molecular sciences*, 11(10):3623–3648.

Moal, I. H., Dapknas, J., and FernándezRecio, J. (2015a). Inferring the microscopic surface energy of proteinprotein interfaces from mutation data. *Proteins: Structure, Function, and Bioinformatics*, 83(4):640–650.

Moal, I. H., Jiménez-García, B., and Fernández-Recio, J. (2015b). CCharPPI web server: computational characterization of proteinprotein interactions from structure. *Bioinformatics*, 31(1):123–125.

Modi, V. and Dunbrack, R. L. (2016). Assessment of refinement of template-based models in CASP11. *Proteins: Structure, Function, and Bioinformatics*, 84(S1):260–281.

Mosca, R., Céol, A., and Aloy, P. (2013). Interactome3D: adding structural details to protein networks. *Nature methods*, 10(1):47–53.

Moult, J. (2005). A decade of CASP: Progress, bottlenecks and prognosis in protein structure prediction. *Current Opinion in Structural Biology*, 15(3 SPEC. ISS.):285–289.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., and Tramontano, A. (2014). Critical assessment of methods of protein structure prediction (CASP) - round X. *Proteins: Structure, Function and Bioinformatics*, 82(SUPPL.2):1–6.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., and Tramontano, A. (2016). Critical assessment of methods of protein structure prediction: Progress and new directions in round XI. *Proteins: Structure, Function and Bioinformatics*, 84(S1):4–14.

Mullard, A. (2012). Proteinprotein interaction inhibitors get into the groove. *Nature Reviews Drug Discovery*, 11(3):173–175.

Mulligan, L. M. (2014). RET revisited: expanding the oncogenic portfolio. *Nature Reviews Cancer*, 14(3):173–186.

Mulligan, L. M., Kwok, J. B. J., Healey, C. S., Elsdon, M. J., Eng, C., Gardner, E., Love, D. R., Mole, S. E., Moore, J. K., Papi, L., Ponder, M. A., Telenius, H., Tunnacliffe, A., and Ponder, B. A. J. (1993). Germ-line mutations of the RET proto-oncogene in multiple endocrine neoplasia type 2A. *Nature*, 363(6428):458–460.

Murray, C. D. and Dermott, S. F. (1999). *Solar system dynamics*. Cambridge University Press.

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., and Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, 2(1):1.

Netzer, Y. and Wang, T. (2011). Reading digits in natural images with unsupervised feature learning. In *Nips*, pages 1–9.

Neveu, E., Ritchie, D. W., Popov, P., and Grudinin, S. (2016). PEPSI-Dock: A detailed data-driven protein-protein interaction potential accelerated by polar Fourier correlation. *Bioinformatics*, 32(17):i693–i701.

Nguyen, H., Roe, D. R., and Simmerling, C. (2013). Improved generalized born solvent model parameters for protein simulations. *Journal of Chemical Theory and Computation*, 9(4):2020–2034.

Nooren, I. M. A. and Thornton, J. M. (2003). Structural characterisation and functional significance of transient proteinprotein interactions. *Journal of molecular biology*, 325(5):991–1018.

Normanno, N., De Luca, A., Bianco, C., Strizzi, L., Mancino, M., Maiello, M. R., Carotenuto, A., De Feo, G., Caponigro, F., and Salomon, D. S. (2006). Epidermal growth factor receptor (EGFR) signaling in cancer. *Gene*, 366(1):2–16.

Nugent, T., Cozzetto, D., and Jones, D. T. (2014). Evaluation of predictions in the CASP10 model refinement category. *Proteins: Structure, Function and Bioinformatics*, 82(SUPPL.2):98–111.

Oliva, R., Vangone, A., and Cavallo, L. (2013). Ranking multiple docking solutions based on the conservation of interresidue contacts. *Proteins: Structure, Function, and Bioinformatics*, 81(9):1571–1584.

Ovchinnikov, S., Kamisetty, H., and Baker, D. (2014). Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *eLife*, 2014(3):e02030.

Ovchinnikov, S., Park, H., Varghese, N., Huang, P.-S., Pavlopoulos, G. A., Kim, D. E., Kamisetty, H., Kyrpides, N. C., and Baker, D. (2017). Protein structure determination using metagenome sequence data. *Science*, 355(6322):294–298.

Pachnis, V., Mankoo, B., and Costantini, F. (1993). Expression of the c-ret proto-oncogene during mouse embryogenesis. *Development*, 119:1005–1017.

Paratcha, G. and Ledda, F. (2008). GDNF and GFRα: a versatile molecular complex for developing neurons. *Trends in Neurosciences*, 31(8):384–391.

Park, H., Lee, H., and Seok, C. (2015). High-resolution proteinprotein docking by global optimization: recent advances and future challenges. *Current opinion in structural biology*, 35:24–31.

Pascanu, R., Mikolov, T., and Bengio, Y. (2012). On the difficulty of training recurrent neural networks. In *Proceedings of The 30th International Conference on Machine Learning*, number 2, pages 1310–1318. arXiv:1211.5063v2.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, arXiv:1201.0490.

Perkins, J. R., Diboun, I., Dessailly, B. H., Lees, J. G., and Orengo, C. (2010). Transient Protein-Protein Interactions: Structural, Functional, and Network Properties. *Structure*, 18(10):1233–1243.

Petrov, D., Margreitter, C., Grandits, M., Oostenbrink, C., and Zagrovic, B. (2013). A Systematic Framework for Molecular Dynamics Simulations of Protein Post-Translational Modifications. *PLoS Computational Biology*, 9(7):e1003154.

Pfeiffenberger, E., Chaleil, R. A. G., Moal, I. H., and Bates, P. A. (2017). A machine learning approach for ranking clusters of docked protein-protein complexes by pairwise cluster comparison. *Proteins: Structure, Function and Bioinformatics*, 85(3):528–543.

Pierce, B. and Weng, Z. (2007). ZRANK: Reranking protein docking predictions with an optimized energy function. *Proteins: Structure, Function and Genetics*, 67(4):1078–1086, 0605018.

Pierce, B. and Weng, Z. (2008). A combination of rescoring and refinement significantly improves protein docking performance. *Proteins: Structure, Function and Genetics*, 72(1):270–279.

Pierchala, B. A., Milbrandt, J., and Johnson, E. M. (2006). Glial Cell Line-Derived Neurotrophic Factor-Dependent Recruitment of Ret into Lipid Rafts Enhances Signaling by Partitioning Ret from Proteasome-Dependent Degradation. *Journal of Neuroscience*, 26(10):2777–2787.

Plaza-Menacho, I., Barnouin, K., Goodman, K., Martínez-Torres, R. J., Borg, A., Murray-Rust, J., Mouilleron, S., Knowles, P., and McDonald, N. Q. (2014). Oncogenic RET kinase domain mutations perturb the autophosphorylation trajectory by enhancing substrate presentation in trans. *Molecular Cell*, 53(5):738–751.

Plaza-Menacho, I., Burzynski, G. M., de Groot, J. W., Eggen, B. J., and Hofstra, R. M. (2006). Current concepts in RET-related genetics, signaling and therapeutics. *Trends in Genetics*, 22(11):627–636.

Pokarowski, P., Kloczkowski, A., Jernigan, R. L., Kothari, N. S., Pokarowska, M., and Kolinski, A. (2005). Inferring ideal amino acid interaction forms from statistical protein contact potentials. *Proteins: Structure, Function and Genetics*, 59(1):49–57.

Pons, C., Talavera, D., de la Cruz, X., Orozco, M., and Fernandez-Recio, J. (2011). Scoring by Intermolecular Pairwise Propensities of Exposed Residues (SIPPER):

A New Efficient Potential for Protein Protein Docking. *Journal of chemical information and modeling*, 51(2):370–377.

Popsueva, A., Poteryaev, D., Arighi, E., Meng, X., Angers-Loustau, A., Kaplan, D., Saarma, M., and Sariola, H. (2003). GDNF promotes tubulogenesis of GFRα1-expressing MDCK cells by Src-mediated phosphorylation of Met receptor tyrosine kinase. *Journal of Cell Biology*, 161(1):119–129.

Pronk, S., Páll, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., Shirts, M. R., Smith, J. C., Kasson, P. M., and van der Spoel, D. (2013). GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*.

Qin, S. and Zhou, H. (2013). Using the concept of transient complex for affinity predictions in CAPRI rounds 2027 and beyond. *Proteins: Structure, Function, and Bioinformatics*, 81(12):2229–2236.

Raina, R., Madhavan, A., and Ng, A. Y. (2009). Large-scale deep unsupervised learning using graphics processors. In *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, pages 1–8.

Raina, R., Ng, A., and Koller, D. (2006). Constructing informative priors using transfer learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 713–720.

Rajgaria, R., McAllister, S. R., and Floudas, C. A. (2006). A novel high resolution CA-CA distance dependent force field based on a high quality decoy set. *Proteins: Structure, Function and Genetics*, 65(3):726–741.

Rajgaria, R., McAllister, S. R., and Floudas, C. A. (2008). Distance dependent centroid to centroid force fields using high resolution decoys. *Proteins: Structure, Function and Genetics*, 70(3):950–970.

Rapaport, D. C. (2004). *The Art of Molecular Dynamics Simulation*. Cambridge University Press.

Rasmussen, S. G. F., Choi, H.-J., Fung, J. J., Pardon, E., Casarosa, P., Chae, P. S., DeVree, B. T., Rosenbaum, D. M., Thian, F. S., Kobilka, T. S., Schnapp, A., Konetzki, I., Sunahara, R. K., Gellman, S. H., Pautsch, A., Steyaert, J., Weis, W. I., and Kobilka, B. K. (2011). Structure of a nanobody-stabilized active state of the β2 adrenoceptor. *Nature*, 469(7329):175–180.

Raval, A., Piana, S., Eastwood, M. P., Dror, R. O., and Shaw, D. E. (2012). Refinement of protein structure homology models via long, all-atom molecular dynamics simulations. *Proteins: Structure, Function and Bioinformatics*, 80(8):2071–2079.

Ravikant, D. V. S. and Elber, R. (2010). Pieefficient filters and coarse grained potentials for unbound proteinprotein docking. *Proteins: Structure, Function, and Bioinformatics*, 78(2):400–419.

Ren, J., Esnouf, R. M., Hopkins, A. L., Jones, E. Y., Kirby, I., Keeling, J., Ross, C. K., Larder, B. A., Stuart, D. I., and Stammers, D. K. (1998). 3'-Azido-3'-deoxythymidine drug resistance mutations in HIV-1 reverse transcriptase can induce long range conformational changes. *Proceedings of the National Academy of Sciences of the United States of America*, 95(16):9518–9523.

Repetto, E., Yoon, I. S., Zheng, H., and Kang, D. E. (2007). Presenilin 1 regulates epidermal growth factor receptor turnover and signaling in the endosomal-lysosomal pathway. *Journal of Biological Chemistry*, 282(43):31504–31516.

Robinson, T. and Fallside, F. (1991). A recurrent error propagation network speech recognition system. *Computer Speech and Language*, 5(3):259–274.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, arXiv:1112.6209.

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Engineering, Design and Selection*, 12(2):85–94.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). *Learining Internal Representations by Error Propagation*. arXiv:1011.1669v3.

Rusmini, M., Griseri, P., Lantieri, F., Matera, I., Hudspeth, K. L., Roberto, A., Mikulak, J., Avanzini, S., Rossi, V., Mattioli, G., Jasonni, V., Ravazzolo, R., Pavan, W. J., Pini-Prato, A., Ceccherini, I., and Mavilio, D. (2013). Induction of RET Dependent and Independent Pro-Inflammatory Programs in Human Peripheral Blood Mononuclear Cells from Hirschsprung Patients. *PLoS ONE*, 8(3).

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, arXiv:1409.0575.

Salakhutdinov, R. and Hinton, G. (2009). Deep Boltzmann Machines. In *AISTATS*. arXiv:1203.4416.

Saul, L. K., Jaakkola, T., and Jordan, M. I. (1996). Mean Field Theory for Sigmoid Belief Networks. *Journal of Artificial Intelligence Research*, 4:61–76, 9603102.

Schalm, S. S., Ballif, B. A., Buchanan, S. M., Phillips, G. R., and Maniatis, T. (2010). Phosphorylation of protocadherin proteins by the receptor tyrosine kinase Ret. *Proceedings of the National Academy of Sciences*, 107(31):13894–13899, arXiv:1604.05974v2.

Schlick, T. (2010). *Molecular Modeling and Simulation: An Interdisciplinary Guide*, volume 21. Springer, 2 edition.

Schmid, N., Eichenberger, A. P., Choutko, A., Riniker, S., Winger, M., Mark, A. E., and Van Gunsteren, W. F. (2011). Definition and testing of the GROMOS force-field versions 54A7 and 54B7. *European Biophysics Journal*, 40(7):843–856.

Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., and Wolfson, H. J. (2005). PatchDock and SymmDock: Servers for rigid and symmetric docking. *Nucleic Acids Research*, 33(SUPPL. 2):W363–W367.

Schuchardt, A., D'Agati, V., Larsson-Blomberg, L., Costantini, F., and Pachnis, V. (1994). Defects in the kidney and enteric nervous system of mice lacking the tyrosine kinase receptor Ret. *Nature*, 367(6461):380–383.

Seemayer, S., Gruber, M., and Soeding, J. (2014). CCMpred - Fast and precise prediction of protein residue-residue contacts from correlated mutations. *Bioinformatics*, 30(21):3128–3130, arXiv:1404.6684.

Segouffin-Cariou, C. and Billaud, M. (2000). Transforming ability of MEN2A-RET requires activation of the phosphatidylinositol 3-kinase/AKT signaling pathway. *Journal of Biological Chemistry*, 275(5):3568–3576.

Shen, M.-Y. and Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. *Protein Science*, 15(11):2507–2524.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., and Lanctot, M. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., and Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature*, 550(7676):354–359.

Simons, K. T., Kooperberg, C., Huang, E., and Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *Journal of Molecular Biology*, 268(1):209–225.

Sivanesan, D., Rajnarayanan, R. V., Doherty, J., and Pattabiraman, N. (2005). In-silico screening using flexible ligand binding pockets: A molecular dynamics-based approach. *Journal of Computer-Aided Molecular Design*, 19(4):213–228.

Škubník, K., Nováček, J., Füzik, T., Pidal, A., Paxton, R. J., and Plevka, P. (2017). Structure of deformed wing virus, a major honey bee pathogen. *Proceedings of the National Academy of Sciences*, 114(12):3210–3215.

Söding, J. (2017). Big-data approaches to protein structure prediction. *Science*, 355(6322):248–249.

Songyang, Z., Carraway, K. L., Eck, M. J., Harrison, S. C., Feldman, R. A., Mohammadi, M., Schlessinger, J., Hubbard, S. R., Smith, D. P., Eng, C., Lorenzo, M. J., Ponder, B. A. J., Mayer, B. J., and Cantley, L. C. (1995). Catalytic specificity of protein-tyrosine kinases is critical for selective signalling. *Nature*, 373(6514):536–539.

Su, X., Ma, J., Wei, X., Cao, P., Zhu, D., Chang, W., Liu, Z., Zhang, X., and Li, M. (2017). Structure and assembly mechanism of plant C2S2M2-type PSII-LHCII supercomplex. *Science*, 357(6353):815–820.

Sutto, L. and Gervasio, F. L. (2013). Effects of oncogenic mutations on the conformational free-energy landscape of EGFR kinase. *Proceedings of the National Academy of Sciences of the United States of America*, 110(26):10616–21, arXiv:1408.1149.

Sutton, R. and Barto, A. (1998). *Reinforcement Learning: An Introduction*, volume 9. arXiv:1603.02199.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. arXiv:1409.4842.

Takahashi, M., Buma, Y., Iwamoto, T., Inaguma, Y., Ikeda, H., and Hiai, H. (1988). Cloning and expression of the ret proto-oncogene encoding a tyrosine kinase with two potential transmembrane domains. *Oncogene*, 3:571–578.

Tansey, M. G., Baloh, R. H., Milbrandt, J., and Johnson, E. M. (2000). GFRalpha-mediated localization of RET to lipid rafts is required for effective downstream signaling, differentiation, and neuronal survival. *Neuron*, 25:611–623.

Terashi, G., Takeda-Shitaka, M., Kanou, K., Iwadate, M., Takaya, D., and Umeyama, H. (2007). The SKE-DOCK server and human teams based on a combined method of shape complementarity and free energy estimation. *Proteins: Structure, Function and Genetics*, 69(4):866–872.

Tobi, D. (2010). Designing coarse grained-and atom based-potentials for protein-protein docking. *BMC structural biology*, 10(1):1.

Tobi, D. and Bahar, I. (2006). Optimal design of protein docking potentials: efficiency and limitations. *Proteins: Structure, Function, and Bioinformatics*, 62(4):970–981.

Torchala, M., Moal, I. H., Chaleil, R. A. G., Fernandez-Recio, J., and Bates, P. A. (2013). SwarmDock: a server for flexible proteinprotein docking. *Bioinformatics*, 29(6):807–809.

Tovchigrechko, A. and Vakser, I. A. (2006). GRAMM-X public web server for protein-protein docking. *Nucleic Acids Research*, 34(WEB. SERV. ISS.):W310–W314.

Tribello, G. A., Bonomi, M., Branduardi, D., Camilloni, C., and Bussi, G. (2014). PLUMED 2: New feathers for an old bird. *Computer Physics Communications*, 185(2):604–613, arXiv:1310.0980.

Tsuzuki, T., Takahashi, M., Asai, N., Iwashita, T., Matsuyama, M., and Asai, J. (1995). Spatial and temporal expression of the ret proto-oncogene product in embryonic, infant and adult rat tissues. *Oncogene*, 10(1):191–8.

Tufro, A., Teichman, J., Banu, N., and Villegas, G. (2007). Crosstalk between VEGF-A/VEGFR2 and GDNF/RET signaling pathways. *Biochemical and Biophysical Research Communications*, 358(2):410–416.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236):433–460.

Uziela, K., Wallner, B., and Elofsson, A. (2016). ProQ3: Improved model quality assessments using Rosetta energy terms. *Nature Publishing Group*, (August):1–10, arXiv:1602.05832.

Vajda, S., Hall, D. R., and Kozakov, D. (2013). Sampling and scoring: A marriage made in heaven. *Proteins: Structure, Function, and Bioinformatics*, 81(11):1874–1884.

Vangone, A. and Bonvin, A. M. J. J. (2015). Contacts-based prediction of binding affinity in proteinprotein complexes. *Elife*, 4:e07454.

Vargas-Leal, V., Bruno, R., Derfuss, T., Krumbholz, M., Hohlfeld, R., and Meinl, E. (2005). Expression and Function of Glial Cell Line-Derived Neurotrophic Factor Family Ligands and Their Receptors on Human Immune Cells. *The Journal of Immunology*, 175(4):2301–2308.

Verga, U., Fugazzola, L., Cambiaghi, S., Pritelli, C., Alessi, E., Cortelazzi, D., Gangi, E., and Beck-Peccoz, P. (2003). Frequent association between MEN 2A and cutaneous lichen amyloidosis. *Clinical Endocrinology*, 59(2):156–161.

Vigneron, N., Stroobant, V., Chapiro, J., Ooms, A., Degiovanni, G., Morel, S., van der Bruggen, P., Boon, T., and Van den Eynde, B. J. (2004). An Antigenic Peptide Produced by Peptide Splicing in the Proteasome. *Science*, 304(5670):587–590.

Viswanath, S., Ravikant, D. V. S., and Elber, R. (2013). Improving ranking of models for protein complexes with side chain modeling and atomic potentials. *Proteins: Structure, Function, and Bioinformatics*, 81(4):592–606.

Vreven, T., Hwang, H., Pierce, B. G., and Weng, Z. (2012). Prediction of protein-protein binding free energies. *Protein Science*, 21(3):396–404.

Wang, R. Y.-R., Kudryashev, M., Li, X., Egelman, E. H., Basler, M., Cheng, Y., Baker, D., and DiMaio, F. (2015). De novo protein structure determination from near-atomic-resolution cryo-EM maps. *Nature methods*, 12(4):335–8.

Wang, S., Sun, S., Li, Z., Zhang, R., and Xu, J. (2017a). Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLoS Computational Biology*, 13(1):e1005324, arXiv:1609.00680.

Wang, S., Sun, S., and Xu, J. (2017b). Analysis of deep learning methods for blind protein contact prediction in CASP12. *Proteins: Structure, Function and Bioinformatics*.

Wang, X. (2013). Structural studies of GDNF family ligands with their receptors - Insights into ligand recognition and activation of receptor tyrosine kinase RET. *Biochimica et Biophysica Acta - Proteins and Proteomics*, 1834(10):2205–2212.

Widrow, B. and Hoff, M. (1960). Adaptive switching circuits. *1960 IRE WESCON Convention Record*, (4):96 – 104.

Wilen, C. B., Tilton, J. C., and Doms, R. W. (2012). Molecular mechanisms of HIV entry. *Advances in Experimental Medicine and Biology*, 726:223–242.

Xiang, Z., Soto, C. S., and Honig, B. (2002). Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction. *Proceedings of the National Academy of Sciences of the United States of America*, 99:7432–7437.

Xu, G., Ma, T., Zang, T., Sun, W., Wang, Q., and Ma, J. (2017). OPUS-DOSP: A Distance- and Orientation-dependent All-atom Potential Derived from Side-chain Packing. *Journal of Molecular Biology*, 429(20):3113–3120.

Yang, Y. and Zhou, Y. (2008). Specific interactions for ab initio folding of protein terminal regions with secondary structures. *Proteins: Structure, Function and Genetics*, 72(2):793–803.

Zacharias, M. (2003). Protein-protein docking with a reduced protein model accounting for side-chain flexibility. *Protein science: a publication of the Protein Society*, 12(6):1271–1282.

Zeiler, M. and Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. In *European conference on computer vision*, pages 1–11. arXiv:1311.2901v3.

Zhang, J. and Zhang, Y. (2010). A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. *PLoS ONE*, 5(10):e15386.

Zhou, H. and Skolnick, J. (2011). GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. *Biophysical journal*, 101(8):2043–52.